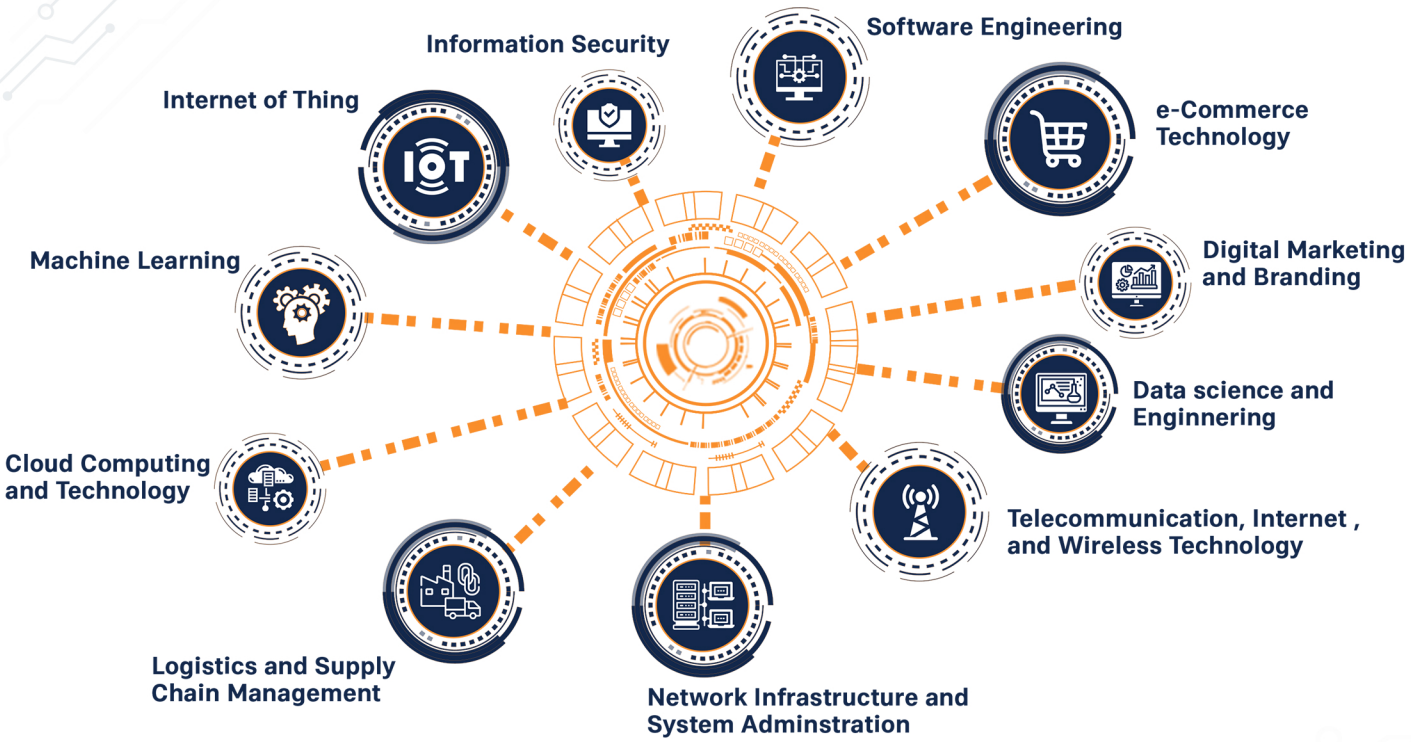
CADT
IDT វិទ្យាស្ថានបច្ចេកវិទ្យាឌីជីថល
Institute of Digital Technology

THE 1ST STUDENT CONFERENCE ON DIGITAL TECHNOLOGY 2022

TOPICS OF INTEREST



DECEMBER 2022

THE 1ST SCDT 2022

LOCAL ORGANISING COMMITEE

CONFERENCE CHAIR

H.E Mr. HEAN SAMBOEUN

MEMBERS

Dr. Srun Sovila
Mr. Chhit Sokna
Mr. Try Vitou
Mr. Keo Saly

CONFERENCE COMMITTEE 2022

H.E Mr. Hean Samboeun
Dr. Srun Sovila
Dr. Khim Chomroeun
Dr. Kong Puthphalla
Mr. Kor Sokchea
Ms. Cheat Morokot
Mr. Chea Socheat
Mr. Teng Chanto
Mr. Nhem Thayheng

Table of Contents

LTE Drive Test for Single Site Verification

Phanny Ath[#1], Morokot Cheat[*2]

[#]*Department of Telecoms & Networking, Cambodia Academy of Digital Technology*

Phnom Penh, Cambodia

[1]phanny.ath@student.cadt.edu.kh

**Ministry of Post and Telecommunications*

Phnom Penh, Cambodia

[2]cheat.morokot@gmail.com

***Abstract*—Communication technology is an important role in the development of all fields and helps many people working in the telecommunications sector. In this context, drive test becomes a component that facilitates service quality testing to achieve the specific coverage that the target desires and to extend signal coverage which is the requirement for a better user experience. Moreover, due to the rapidly increasing number of mobile subscriptions, some areas of eNodeB's cell reached their maximum usage capacity. Hence, the decision to LTE site installation is a best solution for providing good quality of service and better coverage. LTE drive test is a job designed to monitor the use of customer data wherever there is an on-air site. In this paper, we will describe a scenario to conduct drive test and key performance indication for analysis to check the quality of 4G. The main goal is to evaluate and enhance the network performance of a single site where this drive test's pilot area is conducted in the southern part of Cambodia with frequency 2300 MHz in band number 40. A vehicle mount with drive-test equipment and phone set will be used for data collection. PHU Tester is a software used for data collection systems. Google Earth checks on the place to make sure before drive test around site. Genex Assistance is used to analyze log files received from PHU Tester when drive test ends. After the drive test, log files will be analyzed to check on the parameters of RSRP, RSRQ, PCI, Throughput Down Link, and Throughput Up Link. These parameter analyses can provide clear insight for RF engineers to identify and fix primary network issues.**

***Keywords*—Drive Test, 4G, Reference Signal Received Power, Reference Signal Received Quality, Physical Cell Id.**

I. INTRODUCTION

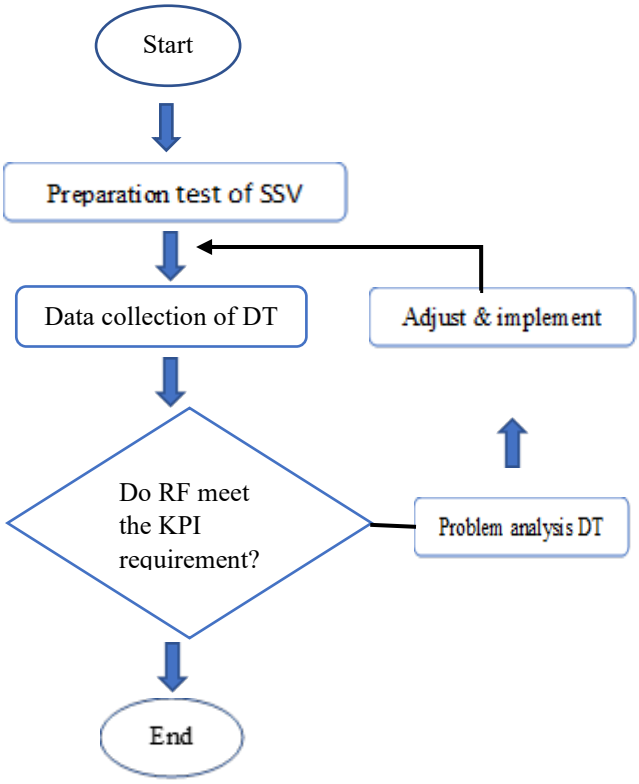
Telecommunication is a solution for far communication that used radio frequency. 4G mobile internet access starts from the first version in release 8 and continues to evolve until release 10, the latest version of LTE-advanced. Its higher data speeds and low-bandwidth radio access technology and packages support flexible bandwidth deployment. could make smartphones much more comparable to PCs, giving them better multimedia and gaming capabilities with speed support a downlink top rate of 326Mb/s and an uplink top rate of 86Mb/s at 20Mhz bandwidth. As mentioned above, we observe that due to the increase in user usage, there is less user support and less coverage, which led to this Drive test. Drive test is to monitor and evaluate the mobile phone network for measuring and capturing the call, drive test around the site and reach location with GPS to get the log file received from the antenna and analyze the cause or the root of the problem to find show solutions [1]. Drive test reflects the RF network as it is important to analyze the data and identify the source of the location channel associated with the GPS system, time and antenna distance to accurately measure [2]. Phone-based drive test for network, coverage assessment provides coverage used for planning, optimization, user complaints, maintenance, and high-rise building cost of drive test. The operator pays for the equipment and staff [3], [4]. To evaluate LTE using drive test, we need to analyze parameters including RSRP for measuring LTE cells over DL. The main purpose of RSRP is to determine the optimal best cell on the DL radio interface and select this cell as the server and results will be sent by UE. With this parameter, different cells using the same carrier frequency can be compared and handover or cell reselection decisions can be taken, in dBm. [5] and [6]. RSRQ is used to determine the best cell for LTE radio connection at a certain geographic location and can be used as the criterion for initial cell selection or handover. signal of cells is received that compare other cells by interference, in dB. [5] and [6]. PCI is used as a cell search process to identify multiple cell transmissions on the same carrier, and the first function aims to perform self-organizing and optimization functions. This is important for reducing the amount of pre-planning and provisioning that operators have to make for LTE networks and for reducing the re-planning required when additional eNBs are deployed. PCI is used in LTE networks as a means for mobile devices to divide different cells. Only ID 504 is used, and at a nearby base station one more special option is used and cells look like for processing configure automatically.
[6] and [4].

This paper purpose presents the log file using the PHU tester tool. Drive test results to show that the site has already been tested, there are some changes before testing. Driving tests depend on the actual location, geography and climatic factors to determine if the results are recognized by optimizing team and the desired goals.

II. METHODOLOGY

A. Main Concept

Drive test was conducted in this research to improve data on service quality. Before driving test, preparation for data collection and path is very important. First, check if our test drive equipment is adequate devices including SIM cards, apps, road maps, and any essentials before leaving the office. Second, when arriving around the site, we start connecting devices to get things started.MS or receiver relative to the transmitter must be included in the test drive. Drive test system consists of GPS, laptop with phone drive test software. GPS and phone are connected to the laptop. Disk tester installed on a laptop. Drive test app to provide an interface between phone and map and start driving around designated sites and get logfiles.



B. 4G Drive Test Analysis Parameters

Reference Signal Received Power: RSRP is key measures of signal level for modern LTE networks. In cellular networks, when a mobile moves from cell to cell and performs cell selection/reselection and handover, it has to measure the signal strength of the neighbor cells. RSRP is defined as the linear average over the power contributions of the resource elements (REs) that carry cell-specific reference signals within the considered measurement frequency bandwidth. signal excellent RSRP value greater than or equals -80dBm dark green as a symbol, dark green as a symbol, good signal value -80 to -90dBm yellow as a symbol, fair signal value -90 to -100dBm orange as a symbol, poor signal value less than equals -100dBm.

$$RSRP\ (dBm) = RSSI\ (dBm) - 10*log\ (12*N) \quad (1)$$

Reference Signal Received Quality: RSRQ are key measures of quality for modern LTE networks. In cellular networks, when a mobile moves from cell to cell and performs cell selection/reselection and handover, it has to measure the quality of the neighbor cells. signal strong excellent RSRP value greater than or equal -10dB dark green as a symbol, good signal value -10 to -15dB yellow as a symbol, fair signal value -15 to -20dB orange as a symbol, poor signal value less than equals -20dB.

$$RSRQ\ (dB) = RSRP\ (dBm) - RSSI\ (dBm) \quad (2)$$

Table 1: Level of Signal Strength of RSRP and RSRQ

RSRP (dBm)	RSRQ (dB)	Signal Strength
>=-80	>=-10	Excellent
-80 to -90	-10 to -15	Good
-90 to 100	-15 to -20	Fair
<=-100	<-20	Poor

Physical cell identity determines the cell ID group and cell ID sectors. There are 168 possible cell ID groups and 3 possible cell ID sectors, so 3*168=504 possible PCI. When cell ID is set to auto, the demodulator will automatically detect cell ID. When cell ID is set to manual, the PHY-layer cell ID must be specified for successful demodulation. Its range is from 0-503 are limited to 504.

$$Cell\ ID = 3*(Cell\ ID\ Group) + Cell\ ID\ Sector \quad (3)$$

C. Flow of Diagram Drive Test

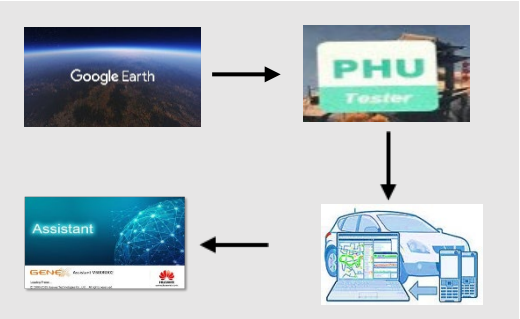


Fig. 1 Flow of Drive Test

i) Google Earth

Google Earth is a geo-browser that accesses satellite and aerial imagery, topography, ocean bathymetry, and other geographic data over the internet to represent the Earth as a three-dimensional globe. Use road maps before going for a

driving test to make sure of latitude and longitude locations and change lanes when an emergency occurs.

Devices for Drive Testing including laptops for immediate logging and analysis, vehicle access to site and GPS phone for navigation. Map where we set the route around the site.

ii) PHU Tester

Step for before Drive Test

- First On Phone: Open the phone, and click on PHU Tester. First, Need login ID, and password, click login, and then settings for set function for uses.

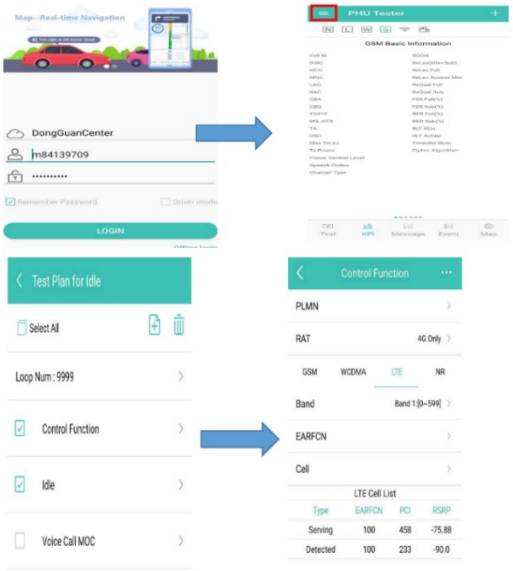


Fig. 2 Step before Drive Test

Then, choose Idle, or DL for testing services 4G, choose Control Faction for set band of 4G which you want, click RAT, choose 4G only, and choose band for testing.

After that, click lock and click running

- On GPS's phone: Open GPS's phone and then you can drive test and analysis already.

iii) Genex Assistance

Create workspaces:

First, Need to Open assistant software goes to new project, create project site name, choose template as default, browse to store data and click ok.

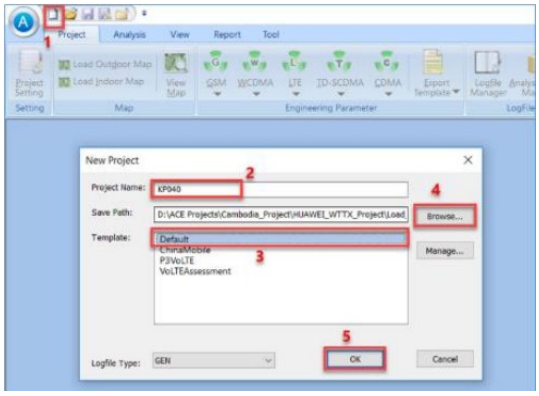


Fig. 3 Create Workspace in Genex Assistance

Second, open project, setting, and choose column LTE,

choose sites to display click PCI, choose others (GCJ-02) and Click ok.

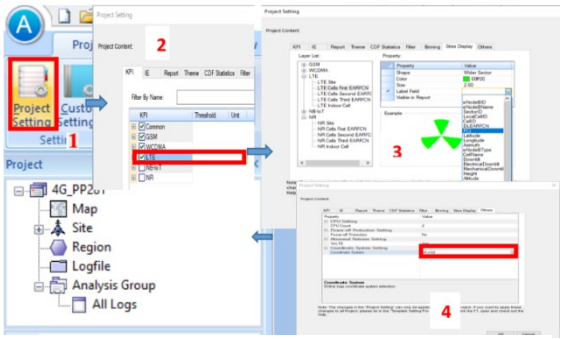


Fig. 4 Click for choosing LTE

Add Engineer parameter

First, open logfile manager, add file, ok, open logfile, and click ok.

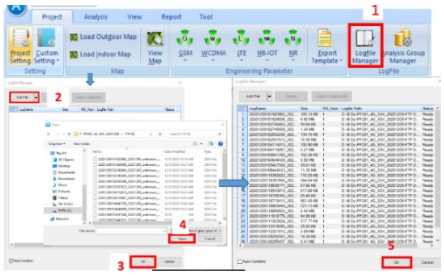


Fig. 5 Add logfile form PC to Genex Assistance

Second, click analysis group manager, add logfile, put name of file, select logfile move to right and click ok.

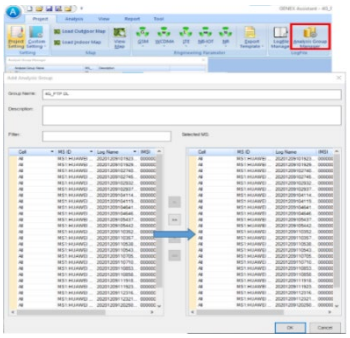


Fig. 6 Add logfile to Analyzed to Right File

Third, all is the processing (running analysis).

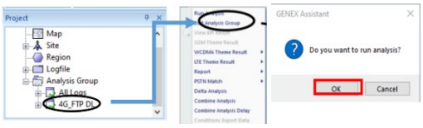


Fig. 7 Choosing which 4G want to Plot Data

Plot Data

First, click LTE, serving and neighboring cells, serving cell, and select PCI, RSRP, etc.

III. ANALYSIS RESULTS

In this Figure 8. the parameter shows the serving Reference Signal Received Power (RSRP), then find a number of samples at display as legend, so look at the map range number of legend from -75 to -40 equals to 38.35% is fair in yellow color and legend -145 to -110 equals 2.29% is very low in red color. With a number to indicate the signal received from the test.

Fig. 8 Serving cell Reference Signal Received Power

In this Figure 9. the parameter shows about Reference Signal Received Quality (RSRQ) of the signal, range number of legends that shows with map, range number of legend from -12 to -10 equals 45.73% that which mean very good by color green and legend -40 to -16 equals 1.93% is very low in red color.

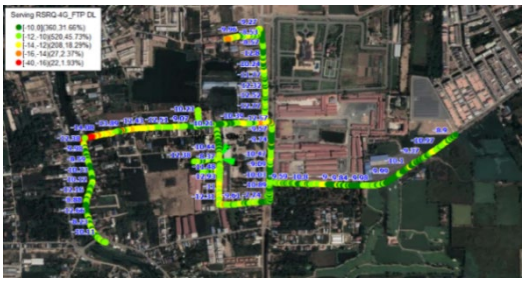


Fig. 9 Serving cell Reference Signal Received Quality

In this Figure 10. PCI plot of site which explains how cells of each site distribute signal, identify distance of site coverage, site cross, or site down.

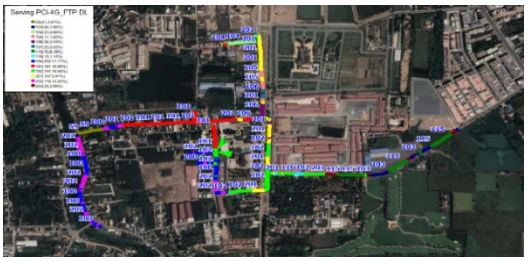


Fig. 10 Physical Cell ID

IV. CONCLUSIONS

In conclusion, LTE Drive Test is an effective method for measuring coverage and quality of the cellular network signal, mobile network latency, voice/data services, user QoE and evaluate the performance KPIs using cellular drive test equipment. The hardware PHU tester equipment and software Genex Assistance and Google Earth are used to analysis main parameters such as power signal quality, power throughput upload and download to analyze the quality of the received signal at that area of the LTE site. If the power received, throughput upload and download of the transmitter are not very good, it is necessary to add cells to areas where coverage does not reach. For next research, method of drive test with multiple site tests with additional parameters and different locations should be considered for improving evaluation.

ACKNOWLEDGMENT

I would like to thank my advisor, Mrs. Cheat Morokot. who was unusually busy with her duties, took the time to guide me through my research on my paper and offered some suggestions on how to improve until I can complete the paperwork successfully.

REFERENCES

[1] S. P. Erik Dahlman, *LTE/LTE-Advanced* for Mobile Broadband, Elsevier, 2011.

[2] M. Rumney, *LTE and the Evolution to 4G Wireless*, John Wiley & Sons Ltd, 2013.

[3] C. S. Seppo Hamalainen, *LTE Self-Organising Networks*, John Wiley & Sons Ltd, 2012

[4] X. Z. Xincheng Zhang, *LTE-Advanced Air Interface Technology*, Taylor & Francis Group, 2013 .

[5] M. Sauter, *From GSM to LTE*, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO198SQ, 2011.

[6] K. G. Ralf Kreher, *LTE SIGNALING*, 2. nd, Ed., John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO198SQ, 2016.

Image Compression using Discrete Cosine Transform

Sombathmorhareach Neath[#1], Morokot Cheat[*2]

[#]*Department of Telecoms & Networking, Cambodia Academy of Digital Technology*

Phnom Penh, Cambodia

[1]Sombathmorhareach.neath@student.cadt.edu.kh

**Ministry of Post and Telecommunications*

Phnom Penh, Cambodia

[2]cheat.morokot@gmail.com

***Abstract*— In recent years, the demand of multimedia and image files are increasing fast, causing to insufficient storage of memory device and slow for data transmission. Therefore, compression becomes more and more significant. An Image compression is defined as an application that compresses data in digital images. Digital images are comprised with large amount of information that requires bigger bandwidth. The techniques of image compression can be generally categorized into two types: lossless and lossy technique. It is an application of data compression that encodes the original image with few bits to reduce the redundancy of the image to minimize the size in bytes. This means that an image where adjacent pixels have almost the same values leads to spatial redundancy. In this paper, we are going to explore theoretical and practical studies of image compression as well as reconstruction. The principal goal of the project is to apply the Discrete Cosine Transform (DCT) technique to compress colour RGB images and grayscale images in order to reveal the project's results. The program for implementing and executing the DCT algorithm of image compression is MATLAB. The result shows that this algorithm can reduce image size with average percentage around 43 % with acceptable quality.**

Keywords— Image Compression, Discrete Cosine Transform (DCT), Inverse Discrete Cosine Transform (IDCT), Quantization, MATLAB.

I. INTRODUCTION

With increasing technology, the amount of data transferred and stored is increasing exponentially. Compression is useful as it helps in reducing the use of expensive resources such as transmission bandwidth and hard disk space. Compression is essential for problem solving and improving the workplace environment. Image compression is a subset of data compression that is used to reduce image data. Its progress based on 2 types: Lossless and Lossy method. Lossless compression techniques play main role in kind of algorithm Huffman Coding, Run-Length Coding. Usually, every image in each shape contains a large number of the same blocks of the same colour. In the process of scanning the colour blocks, many consecutive scanning lines or continuous pixels on the same scanning line carry the same colour value when storing the image. Meanwhile, run-length coding is a method that stores a pixel value and the number of pixels with the same colour rather than storing the same colour blocks in the image one by one. So, run-length coding has a high compression rate when an image has a large amount of colour in the same blocks [1]. Huffman coding is a form of statistical coding that attempts to reduce the number of bits required to represent a string of symbols. So, it will reduce the number of bits that it gets from the value of an image pixel by traversing the Huffman tree. Simply, bits number lengths vary and will be shorter for the more frequently used from pixel value. The algorithm was proposed by Dr. David A. Huffman in 1952 [2].

Otherwise, lossy compression algorithms are techniques that reduce file size by discarding the less important information to make sure that a little quality loss is acceptable to achieve a significant bit rate reduction which is called Transform Coding such as Discrete Cosine Transform (DCT), Discrete Fourier Transform (DFT). Transform coding is a method of encoding an image's transform coefficients by utilizing correlation of all pixels in the image. Image data will enter a new transform domain after special transformation due to the strong correlation of image data. In this new coordinate space, the original scattered image data can be centrally distributed. Because most images' transform coefficients are small, compression coding can be accomplished by removing the coefficients that are close to "0." The loss of irrelevant data will be negligible after the image is reconstructed using the inverse transform [1]. DFT is a method that converts a sequence of N complex numbers into another sequence of complex numbers defined by Euler's formula, where the last expression follows from the first [3]. It is a traditional method for converting images from the space domain to the frequency domain by adding another dimension to image observation and image frequency distribution characteristics [4] whereas DCT is a method that transforms images from the spatial domain to the frequency domain while encoding the image using term of sum of cosine functions, dividing it by quantized standard JPEG, and decoding the IDCT formula while transforming the images from the frequency domain back to the spatial domain [5].

According to S.S Pandey et al (2015) and M.N Rasheed et al (2020), the discrete Fourier transform takes longer to compute than the DCT due to its complex algorithm and time-consuming process [3][4]. In the research of R. A. Hamzah et al (2021) with using Discrete Cosine Transform compression technique, users are able to compress and reduce the size of

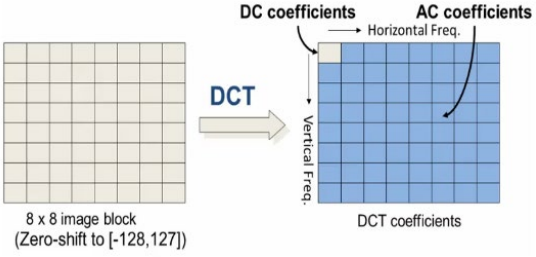
image in order to store the image with small size of capacity under condition saved within JPG format, even those original files are JPG, PNG or BMG [5].

This paper proposes an algorithm of the Discrete Cosine Transform method to compress images and compares results. The process of compression is calculated using own coding in MATLAB. Experiments are used to test colour and grayscale images with extension of PNG and JPEG to determine which one reduces better. After compressing the image, the results will show savings percentages of image size before compression and after compression.

II. THEORETICAL BACKGROUND

A. Discrete Cosine Transform

The Discrete Cosine Transform (DCT) attempts to divide the image into parts; each coefficient is encoded into a separate block. It transforms image from the spatial domain to the frequency domain. Before calculating the DCT of the block, its values are shifted from a positive value to one centred around zero. Because the DCT is designed to work on pixel values ranging from -128 to 127, the original block is "levelled off" by subtracting 128 from each entry [6].



The forward 2-D Discrete Cosine Transform (DCT) of $M \times N$ block of pixels is defined as:

$$F(u,v)=\alpha_u\alpha_v\sum_{j=0}^{M-1}\sum_{k=0}^{N-1}f(j,k)\cos\frac{(2j+1)\pi u}{2M}\cos\frac{(2k+1)\pi v}{2N},\begin{cases}0<u<M-1\\0<v<N-1\end{cases} \quad (1)$$

where $\alpha_u=\begin{cases}\frac{1}{\sqrt{M}},u=0\\\sqrt{\frac{2}{M}},1\le u\le M-1\end{cases}$ $\alpha_v=\begin{cases}\frac{1}{\sqrt{N}},v=0\\\sqrt{\frac{2}{N}},1\le v\le N-1\end{cases}$,

f(j,k) is the intensity of pixel in row *j* and column k,
F(u,v) is the DCT coefficient in row *u* and column v,
M and *N* is size of block of DCT (*M=8, N=8*).

B. Inverse Discrete Cosine Transform

The Inverse Discrete Cosine Transform (IDCT) attempts to reconstruction of image by decoded the bit stream representing the quantized from separate parts. It transforms a frequency domain to image domain. The inverse Discrete Cosine Transform (IDCT) is defined as:

$$f(j,k)=\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}\alpha_u\alpha_v F(u,v)\cos\frac{(2j+1)\pi u}{2M}\cos\frac{(2k+1)\pi v}{2N},\begin{cases}0<u<M-1\\0<v<N-1\end{cases} \quad (2)$$

where $\alpha_u=\begin{cases}\frac{1}{\sqrt{M}},u=0\\\sqrt{\frac{2}{M}},1\le u\le M-1\end{cases}$ $\alpha_v=\begin{cases}\frac{1}{\sqrt{N}},v=0\\\sqrt{\frac{2}{N}},1\le v\le N-1\end{cases}$,

F(u,v) is DCT coefficient represented by the matrix,
N and *M* is the size of the block that the DCT is done on. N and M equals 8 and *u* and *v* range from 0 to 7.

C. Quantization

Quantization is the step where most of the compression takes place. Quantization is achieved by compressing a range of values to a single quantum value. Quantization is achieved by dividing transformed image matrix by the quantization matrix used. Values of the resultant matrix are then rounded off. In the resultant matrix coefficients situated near the upper left corner have lower frequencies. Simply put, it reduces the number bits needed to store an integer value by reducing the precision of integer. A typical quantization matrix, as specified in the original JPEG Standard[7], is as follows:

$$Q(u,v)=\begin{bmatrix}16&11&10&16&24&40&51&61\\12&12&14&19&26&58&60&55\\14&13&16&24&40&57&69&56\\14&17&22&29&51&87&80&62\\18&22&37&56&68&109&103&77\\24&35&55&64&81&104&113&92\\49&64&78&87&103&121&120&101\\72&95&95&98&112&100&103&99\end{bmatrix} \quad (3)$$

where Q matrix is the normalization array, which has been tested empirically and found to generate good results more than another matrix. By dividing each element in the transformed image matrix from DCT by the corresponding element in the quantization matrix, and then rounding to the nearest integer value. Meaning, after calculating the DCT, the next step is to find and discard the coefficients that contribute the least to the image.

The normalized and quantized coefficient is:

$$F^Q(u,v)=NearestInteger\left(\frac{F(u,v)}{Q(u,v)}\right) \quad (4)$$

The inverse quantization, which means simply that the normalization is removed by multiplying by the quantizer matrix, returns the result for input to the IDCT with formula:

$$F^{Q'}(u,v)=F^Q(u,v)\times Q(u,v) \quad (5)$$

D. Saving Percentage

A specific number has been assigned to determine the ideal quality. While percentages are easier to understand when determining the quality of an image after compression.

$$SavingPercentages=\frac{SizeBeforeCompression-SizeAfterCompression}{SizeBeforeCompression}\times 100\% \quad (6)$$

III. METHODOLOGY

A. Main Concept

Discrete Cosine Transform algorithm is applied in this research to reduce image size. The concept of this algorithm is the standard to compressed on JEPG and PNG format. Using the original image to break into 8x8 blocks of pixels, which working from left to right, top to bottom while the DCT is applied to each block. Then each block is compressed through

quantization. The array of compressed blocks that constitute the image is stored in a drastically reduced amount of space. When desired, the image is reconstructed with a process that uses the Inverse Discrete Cosine Transform.

B. Concept of Compression

The structure of the majority of discrete cosine transform compression techniques is generally similar, as seen in Fig. 1. Typically, importing an image file comes first. Commonly, an image file is in the JPG or PNG format. The next step is to organize each image file into 8x8 blocks. To obtain the DCT coefficients, the discrete cosine transform (DCT) tries to apply it to each block. When each DCT coefficient has been quantized (i.e., split by a number and rounded to the next integer), The DCT coefficients are quantized differently because humans struggle to distinguish between differences in high-frequency and low-frequency components. The divisor for the low-frequency DCT coefficients is significantly less than that for the high-frequency DCT coefficients. After quantization, many high frequency components are rounded to zero. The two-dimensional DCT coefficients block is ordered into a one-dimensional coefficient stream using "Zig-Zag" sequence to facilitate entropy coding by placing low frequency coefficients to high frequency coefficients.

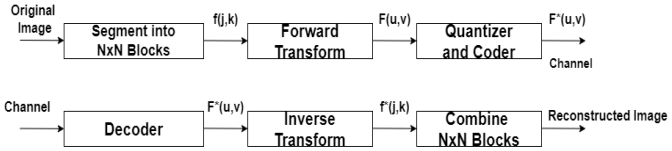


Fig. 1 Overall Concept

C. Flow of Diagram

The flow of diagram is illustrated in Fig. 2.

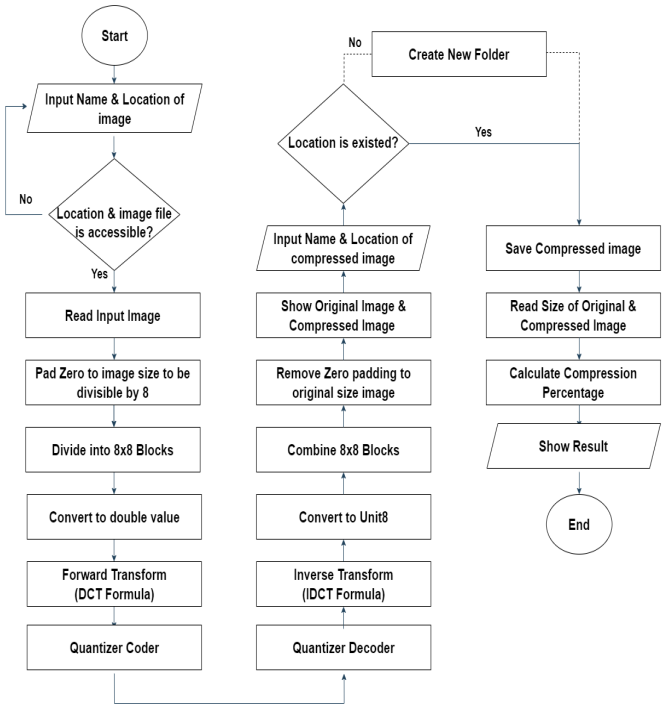


Fig. 2 Flow Process

Firstly, the process will take the image within the location and name of the image that the user has input with the function command, and then read pixel value of the image. After that, image will be divided into sub-images with 8x8 blocks, then execute the forward transform one by one using the DCT formula, continue by quantizing the values that have been transformed to keep only the important values, and finally encode them one by one. After the image has been transformed with DCT, proceed with the decoding, then inverse transform each value one by one using the IDCT formula, and finally combine the small 8x8 blocks together, which is the final step of the process. Finally, all the results will calculate saving percentages with display images and save a compressed image with a name and location.

IV. EXPERIMENTAL SIMULATION AND RESULT

Image compression is an important technique for reducing image size and sending it in such a small size. Images have been developed on various platforms and used different formats or standard. Today's most widely used are PNG format and JPEG format. The experiments were conducted with twelve images which has different shapes and sizes of files and different formats. Each image was in good quality and clear. Table 1 Before experimental process, all images needed to compress by order of the number file which start from 1. in PNG or JPG format to 6. in PNG or JPG format. When each image has been compressed, make a note of the results and save them as compressed image files.

A. Result by Each Process

Fig.3 is result of overall flow from divide whole image into 8x8 block of pixels by converse to double value to process DCT through quantization, and then inverse quantization to input into IDCT; after that, convert value of pixel into uint8 value and finally merge each 8x8 block of pixels together.

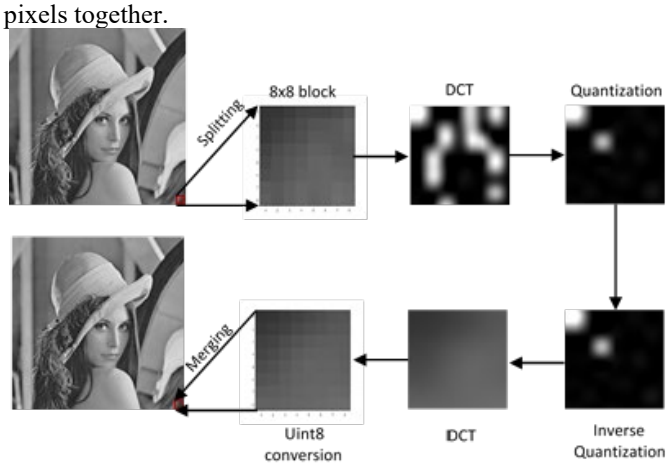


Fig. 3 Overall Result of Each Process

As in result in Fig. 4, each 8x8 block of pixels are combined to visualize for understanding. It has shown that value after DCT transformation and quantization will focus on location of important information for image (high frequency), so it will ignore some useless information for compression.

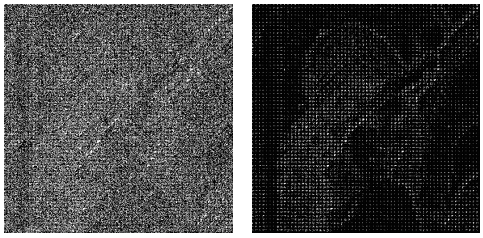


Fig. 4 Block Combination of DCT and Quantization

B. *Overall Compression Result*

Fig. 5(a) and Fig. 5(b) are sample images for experiment, which Fig. 5(c) and Fig. 5(d) are results from compression algorithm. The images after compression, quality of compressed images is acceptable with compressed rate from size 701 KB to 515 KB and 1.72 MB to 953 KB, respectively.

(a) JPG image before compression with size 701 KB

(b) PNG image before compression with size 1.72 MB

(c) JPG image after compression with size 515 KB

(d) PNG image after compression with size 953 KB

Fig. 5 Images before and after Compression

The overall result is in Table 1 to present reduced size in JPG format, which is much better than PNG format, even though each image has a different size and shape. In other words, it means that PNG store with lossless compression algorithm that not compatible to use with the Discrete Cosine Transform algorithm, whereas JPG uses a lossy compression algorithm that is compatible with the Discrete Cosine Transform algorithm.

Table 1 Image Comparison

Comparison			**Save%**
Scale	**Original**	**Compression**	
1.jpg	701 KB	515 KB	26.48%
1.png	4.58 MB	4.27 MB	6.80%
2.jpg	295 KB	120 KB	59.25%
2.png	2.46 MB	1.14 MB	53.33%
3.jpg	610 KB	301 KB	50.65%
3.png	3.15 MB	2.94 MB	6.70%
4.jpg (Grayscale)	525 KB	259 KB	50.68%
4.png (Grayscale)	1.72 MB	953 KB	46.13%
5.jpg	497 KB	181 KB	63.49%
5.png	2.49 MB	1.58 MB	36.55%
6.jpg (Grayscale)	64.0 KB	28.6 KB	55.34%
6.png (Grayscale)	163 KB	111 KB	31.34%

The chart of Fig. 6 illustrates the difference between JPG and PNG is that their sizes are smaller because they reduce the difference between their original and compressed sizes. Besides, the PNG result also shows that the compression ratio for PNG is the lowest. JPG has a higher compression ratio than PNG, as previously observed.

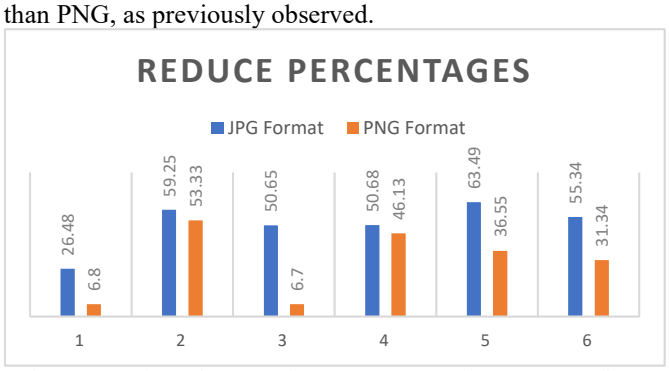


Fig. 6 Comparison of Compression Percentage (%) between JPG and PNG

V. Conclusions

In conclusion, Discrete Cosine Transform (DCT) algorithm, which is a lossy compression technique, can compress image size including color and grayscale image as well as type of PNG and JPG with effective result and acceptable quality. The compression can be on other image format in addition to PNG and JPG in the experiment. For next research, the experiment should test more images with modification of coding which can give better and accuracy results and change from MATLAB experiment into Python experiment.

Acknowledgment

I would like to thank Cambodia Academy of Digital Technology who give me an opportunity to write this paper, and especially to my mentor, lecturer Cheat Morokot who guide me to write this paper from the beginning until we have completed this paper. I would like to pay my respect for her time to pay attention on this paper with me.

References

[1] L. Tan, Y. Zeng, and W. Zhang, "Research on Image Compression Coding Technology," *J. Phys. Conf. Ser.*, vol. 1284, no. 1, 2019, doi: 10.1088/1742-6596/1284/1/012069.

[2] A. A. Shaikh, P. Gadekar, and V. Vikhe, "Huffman Coding Technique for Image Compression," *An Int. J. Adv. Comput. Technol.*, vol. 4, no. 4, pp. 1585–1587, 2015.

[3] M. H. Rasheed, O. M. Salih, M. M. Siddeq, and M. A. Rodrigues, "Image compression based on 2D Discrete Fourier Transform and matrix minimization algorithm," *Array*, vol. 6, p. 100024, 2020, doi: 10.1016/j.array.2020.100024.

[4] S. S. Pandey, M. P. Singh, and V. Pandey, "Image Transformation and Compression using Fourier Transformation," *Int. J. Curr. Eng. Technol.*, vol. 5, no. 2, pp. 1178–1182, 2015.

[5] R. A. Hamzah, M. M. Roslan, A. F. Bin Kadmin, S. F. B. A. Gani, and K. A. A. Aziz, "JPG, PNG and BMP image compression using discrete cosine transform," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 19, no. 3, pp. 1010–1016, Jun. 2021, doi: 10.12928/TELKOMNIKA.v19i3.14758.

[6] K. Cabeen and P. Gent, "Image Compression and the Discrete Cosine Transform," *Coll. Redwoods*, pp. 1–11, 1999, [Online]. Available: papers2://publication/uuid/26D3790F-6B32-4D53-9755-90C1CCBC63F5

[7] A. Baskurt, "Numerical image compression using the discrete cosine transform," *Signal Processing*, vol. 19, no. 4, p. 346, 1990, doi: 10.1016/0165-1684(90)90166-v.

In-Building Solution Using Passive Distributed Antenna System to Improve Cellular Coverage

KORM Sornya[#1], CHEA Socheat[*2]

Department of Telecoms and Networking, Faculty of Digital Engineering,

Institute of Digital Technology, Cambodia Academy of Digital Technology, Cambodia

[#1]sornya.korm@student.cadt.edu.kh

[*2]socheat.chea@cadt.edu.kh

***Abstract*— To avoid some problems of the mobile traffic generated by indoor users of many new high-rise buildings in Cambodia a system is designed to overcome those problems known as IBS. IBS is the technology or solution to apply in the building and distribute the cellular signal of mobile operators. Operators choose to use a Micro BTS site instead of implementing a Macro Site. That signal from a microsite BTS site is distributed through ca-axial lines and DAS. Moreover, DAS is a system to deal with isolated spots of poor coverage inside a large building. DAS can be active, passive, or hybrid based on implement process. In this case at GIA Tower, GIA Tower has implemented passive DAS of IBS to solve all those problems. It also can be the best option to solve all these problems of mobile traffic. A passive DAS System is one of the most popular solutions for IBS, as it captures cellular signals from outside cell towers through the use of an antenna without using active components and no need power supply. Especially, Passive is cheaper than active DAS The idea of this system is to extend and distribute the cellular signal coverage inside the building to be good as outside with better signal coverage, high quality, high data capacity, etc. Additionally, the frequency band has to be perfectly planned to avoid interference from outside coverage as they share the same frequency band.**

***Keywords*— Indoor Building Solution, Distributed Antenna System, Key Performance Indicator, Walk-Test, Signal and Power Source.**

I. INTRODUCTION

Every Building (Hospital, Hotel, Condo, Airport, Department, Supermarket, Mall, etc…) has its challenges that will affect the people while using the internet inside the building [1]. From day to day, technology is up to date and requires good performance with a high-quality network, and always have a new invention to make life easier and better connection with the global. For telecommunication industry has also developed from 1G to 5G to become more efficient [1]. On-air each site, may have many teams to contribute to each other such as planning, installation, optimization, walk test, etc. Even the operators or engineers try to look for the new and fastest technology outside the building but it still needs to update inside the building to cover more performance as well. To address these challenges, an In-building solution is still the best option or technique that will connect outdoor networks allowing better service.

The mobile traffic is generated by indoor users which cause high traffic jam which leads to leakage in coverage in addition to building penetration losses and air losses, so a system is designed to overcome those problems known as IBS. That we have faced problems with mobile coverage as outside towers are not enough to cover the inside of the buildings. As this problem also affects IBS as well by facing some parts of capacity issues, coverage issues, and quality issues. IBS has provided telecom network solutions comprising a distributed antenna system (DAS) enhancing coverage and capacity inside the buildings with weak or no telecom signal [1], [2]. IBS provides indoor coverage using a series of hubs/equipment distributing the signal to several antennas and it obtains signals from diverse sites for this case operators choose to use a Micro BTS site instead of implementing a Macro site. The signal from a microsite BTS site is distributed through co-axial and DAS.

II. RELATED WORK

In-building Solution (IBS) can work and become the most important role that could get the signal inside the building because of the distributed antenna system (DAS). IBS uses DAS technology to build infrastructure [3]. The distributed antenna system concept enhances or fixes the capacity and diversity of next-generation wireless communication networks, due to the inherently added micro and macro diversity [4]. And it clustered the installation of antennas to boost cellular network coverage in areas with a weak signal. Distributed antennas system have been recently shown to provide considerable gains in coverage and capacity, at a lower cost than decreasing cell size [5]. Whichever signal source of a system is used, a DAS needs to amplify, distribute and rebroadcast it throughout the building.

Especially, DAS also deals with isolated spots of poor coverage inside a large building or area by installing a network of relatively small antennas throughout the building to serve as repeaters. Once received, the cellular signal must be distributed throughout the building. That there are three main types of DAS signal distribution systems that can use. Like, Passive, Active, and Hybrid DAS. A distributed antenna system's performance depends on the type of technology it uses and depending upon the components used in the system. Additionally, related to the IBS implementation process, site optimization is also the main part after finishing the

implementation. Signal status verification on each portion of the building. Ensuring proper implementation of DAS as per the design requirements. By doing the walk-test to check the performance of the building.

III. Methodology

A. Main Concept

The building owner wants to get the signal inside the building. To receive the signal through the GIA tower all the engineers must find out and learn more about the building step by step. Especially, to decide on one of the DAS solutions that could apply at GIA Tower. Before choosing or deciding to use one of those techniques or solutions all the engineers have to follow step by step as shown in Figure 1. Because putting or choosing one of those solutions from the step is very important before applying it to the building. In this case, people have to make a note or notice this as well. Doing a site survey and link budget is the most important case of all the steps. That is such an uneasy and hardest point to do during IBS. After installation of DAS, it also needs to do the walk testing and optimizing in purpose to check the cellular coverage signal and quality for a good KPI performance.

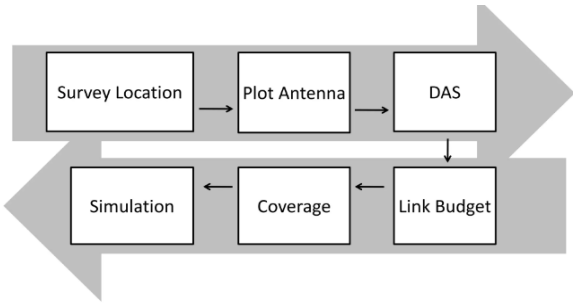


Fig. 1 Plot Diagram Indoor DAS

B. Flow of Diagram

Likewise, GIA Tower IBS implement a process after doing the site survey, and step by step it decides to choose Passive DAS to apply in the building. The Passive DAS is the most widely distributed system for IBS, especially for small buildings or small areas, but sometimes it could be applied in the medium building as well depending on the situation. They are also easy to install and hence in terms of cost they turn out to be the least expensive option than other DAS solutions. It is relatively easy to plan and can be implemented in a harsh environment [6]. Passive DAS captures the signal source (which is BTS outside or Antenna on the roof) and connects to a cellular amplifier or repeater, which then connects to the distributed antenna around the building. Moreover, it provides an amplified re-broadcasted signal from the carrier cell tower, using passive components like coaxial cable, splitters, and diplexers to distribute the signal. But on the other hand, it could suffer from signal loss. A typical passive DAS design is shown in Fig 2.

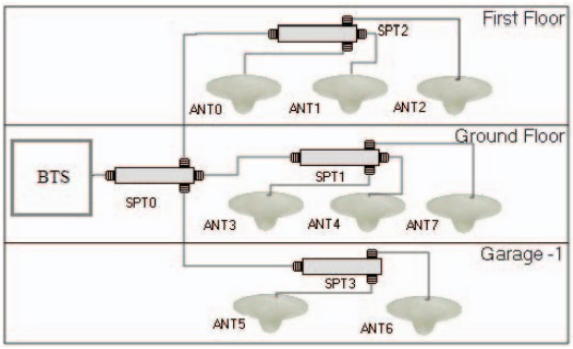


Fig. 2 Sample of Passive DAS, a traditional indoor system composed of the passive element

IV. Implementation

The concept of IBS at GIA Tower has been divided into two parts: Indoor and 3 Outdoor sectors. That the signal source and power system are provided in the IBS room on the 7 floors and 29 floors.

A. Power System

The power source gets from the owner of the building and Edotco (Telecommunication infrastructure service company) as well as build the infrastructure of IBS, such as below:

1) AC source will provide by the owner building from Electricite du Cambodge (EDC).
2) ACPDB or ACDB to provide AC source to UPS
3) An uninterruptable power supply or Uninterruptible power source (UPS) is used to protect hardware, telecommunication equipment, or other electrical equipment where an unexpected power disruption could cause injuries or data loss and also connected to Battery and Rectifier that is used to back up the power into the battery. Typically, DC power with support -48v to equipment.
4) The rectifier is connected to UPS is an electrical device that converts AC power to DC power, which flows in only one direction. And then providing the DC power to Telecom equipment and in the batteries protect when AC power is lost.

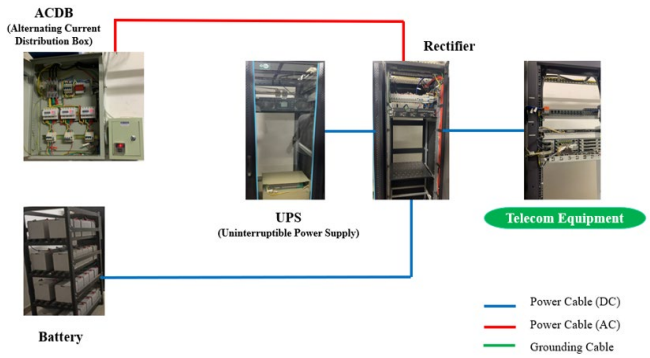


Fig. 3 Flow works of Power Source System

B. Signal Source

The Signal Source and Power get from the DAS system that connects to the IBS room. Elevators access room used for IBS cable running from RF to each floor. As you have shown in Fig 4:

1) The signal source gets from the Passive DAS in the IBS room and the main signal source gets from the nearby site connected to 3 outdoor sectors of BTS by using optical cables that go into the ODF. And then ODF will be transmitting to BBU.
2) The ground cable is mostly copper wire; one end is connected to Passive telecom equipment and another end is connected to Bus Bar nearby them. Thus, it provides the surge and high voltage protection of RRU and keeping safe from natural weather and climate.
3) BBU connected directly to the RRU by using the optical fiber cable.
4) Then, RRU connected to the DAS system by using a Jumper cable.
5) Finally, BBU connected to the RRUs to distribute the cellular signal source to the Omni-Antenna and Directional-Antenna into the building to ensure that it will expand the signal coverage.

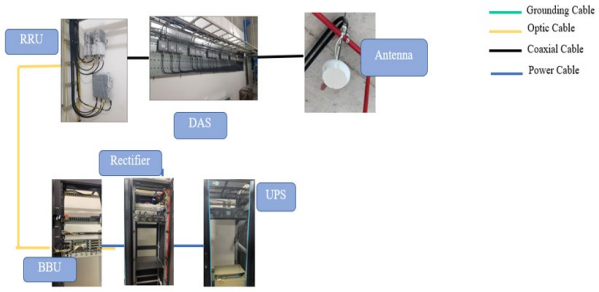


Fig. 4 Signal Source

Especially, here is all the telecom equipment that use in the IBS room with processing in Passive DAS. Embedded Power provides stable -48VDC power for wireless and fixed access networks, transmission networks, and enterprise network equipment. The equipment that has been set up in the 19" open rack is Rectifier, Base Band Unit (BBU), Direct Current Distribution Unit (DCDU), Radio Transmission Network (RTN), and Optical Distribution Frame (ODF).

Fig. 5 Telecom Equipment in 19" open rack

V. Installation And Resting Result

A. Final Installation

Indoor System installation and configuration:

- A signal source is connected to Passive DAS by using a power cable (DC).
- Passive DAS has 2 sectors that are connected to BBU at 7 floors and the other sector connected at 29 floors by using the jumper and optical fiber.
- Passive DAS connected to 2 RRU with the different frequency ranges of the signal by coaxial cable.
- RRU was connected by jumper and splitter to Omni and Panel Antenna to each floor.

Outdoor System installation and configuration:

- ACDB connected to Blade Power outside that connects to 3 RRU outdoor through coaxial cable.
- Blade Power converts AC source to DC to support the telecom Equipment on the high floor.
- RRU connects to BBU inside the room by using fiber optic.
- RRU connects to Panel Antenna by using a Jumper cable.

TABLE I
ANTENNA USING ON BUILDING

floor	ANT Omni/ANT Panel
Car Park, Office, Condo (43F)	289
Lift (43F)	145
Total	434

TABLE II
ANTENNA USING WITH 3 SECTORS

Sector 1	183
Sector 2	123
Sector 3	128

B. KPI Testing Parameters

The last main point of the Passive DAS, after finishing the installation of IBS at GIA Tower we have to do the walk test in the building from the basement floor to the top of the building. A walk test is a tedious and time-consuming task for mobile network operators, typically required to verify signal levels and the quality of service provided by a newly deployed cell site. The resources spent on such tests can be reduced with the use of accurate planning and designing tools [7]. But by doing the whole walk test with the specific information focus only on 3G and 4. By the way, the last walk test is to check the coverage signal and verify with the submission

report too. Here is the walk test result of (IBS) DAS at GIA Tower:

- Sector1 covers from B1F to 7F, and 3G voice quality and 3G/4G throughput are good.
- Sector 2 covers from 8F to 25F, 3G voice quality and 3G/4G throughput is good.
- Sector 3 covers from 26 floors to 43 floors, and 3G voice quality and 3G/4G throughput is good.
- 3G voice no drops Inside lift from ground to top.

Here is some of the requirement for walk test testing some parts of 3G and 4G:

- **RSRP**: Reference Signal Received Power (RSRP) is a measurement of the received power level in an LTE cellular network. The average power is a measurement of the power received from a single reference signal. During measurements, Reference Signal Received Power (RSRP) is calculated per each eNodeB (DL) signal available at a given location [7].
- **RSRQ**: Reference Signal Received Quality. This again only applies to 4G LTE networks and is a measure of the signal quality of a cellular connection. RSRQ is typically displayed in a range from 0dB (highest quality) to -20dB (lowest quality). Typically, better signal quality results in a more reliable connection [7].
- **SINR**: Signal to Interference & Noise Ratio measures signal quality: the strength of the wanted signal compared to the unwanted interference and noise. Mobile network operators seek to maximize SINR at all sites to deliver the best possible customer experience, either by transmitting at higher power or by minimizing interference and noise [7].
- **Throughput**: Throughput is the actual amount of data that is successfully sent/received over the communication link. Throughput is presented as kbps, Mbps, or Gbps, and can differ from bandwidth due to a range of technical issues, including latency, packet loss, jitter, and more. And there sometimes should be DL/UL throughput testing [7].
- **RSCP**: In the UMTS cellular communication system, received signal code power (RSCP) denotes the power measured by a receiver on a particular physical communication channel. It is used as an indication of signal strength, as a handover criterion, in downlink power control, and to calculate path loss.
- **Ec/No and Ec/Io**: relates to signal-to-noise ratio before dispreading (i.e., before rake receiver) Ec/No is the ratio between the received energy from the pilot signal CPICH per chip (Ec) to the noise density (No). In another way, we can tell, that it is the ratio of the received signal level (RSCP) to the sum of all levels of signals on the same frequency (RSSI). Therefore, the higher Ec / No, the better the difference between the signal and noise. Ec/Io stands for Energy per chip to Interference power ratio. It is measured before dispreading in WCDMA systems.

RF Experience	RSRP (dBm) Reference Signal Received Power	RSRQ (dB) Reference Signal Received Quality	SINR (dB) Signal to Interference & Noise Ratio
Excellent	>= -80	>= -10	>= 20
Good	-80 to -90	-10 to -15	13 to 20
Mid Cell	-90 to -100	-15 to -20	0 to 13
Cell Edge	<= -100	<= -20	<= 0

Fig. 6 Signal Strength

C. *Walk Testing Result*

For the final test of the signal performance in the building, we are checking as we mention above. After we do the final analysis of the signal, it's almost fine. Even if it has some loss and interference in the building but it still gives us and other users with good results and good coverage, capacity, and quality as well.

Here are some results of the walk test for the sector1 at B1 floor:

- RSRP: as we see in the figure, RSRP is good at some points that we can see by the legend which means the range from -90 to -80 is excellent in light green color and -100 to -90 is very low in pink color.
- RSRQ: measure the quality of the signal that covers sites. Based on the region, the range from -10 to 0 is excellent in green color and -40 to -20 is very low in pink color.
- SINR: measure the power and how long that site can adjust signal and including noise. The figure shows that the point is far from the site and getting a low signal. Based on legend, the range from 20 is excellent in green color and 0 is very poor in pink color.
- Throughput DL: In the figure, we see throughput DL is so poor that the range from 90000 is good with green color and 0 to 4000 is bad with red color. It is so low; it can be high users.

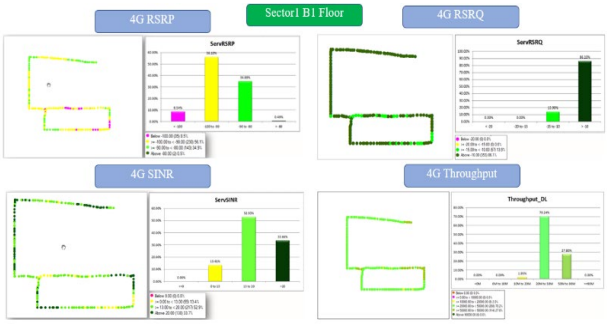


Fig. 7 Final Result of the walk test

Network Cell Information is used to test the performance of the signal as well as to make sure that everything that has been installed and optimized is working well. But there are many tools or software that can use for testing as well. In this

case, we test only one round per floor. It means that if the signal was found and able to accept, we have to test only one time. But on the other hand, if during testing we find out some problems like (loss, interference, noise, no signal, or lower signal in some part) we need to verify and check this case again and start doing the walk test for those floors again.

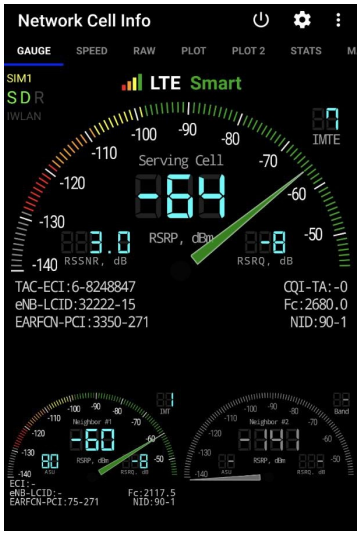


Fig. 8 Network Cell Information

VI. Conclusions

As a result, after implementing the Passive DAS System in the building, it is better coverage than before, people could access the signal in the high place and are much better at using the internet and call service in different mobile operators. So, DAS is more powerful for residence areas, markets or tunnels. Furthermore, I learn and find out so deeply about In- Building Solution Techniques, reflexing with the others on the team, and how to solve problems reliably. From the point of view, Passive DAS solutions can be in a different way as well depending on the problems. And some parts could be a little bit different depending on the situation and style of a problem serving as well.

Acknowledgment

Apart from the efforts, I'd want to convey my heartfelt appreciation and gratitude to CADT, for creating this program. Likewise, I would like to show my appreciation to my advisor **Mr. Chea Socheat** who is working so hard to help, guide, encourage, teach and give me powerful support. When I came across an obstacle. Last but not least, I greatly thank my supervisor **Mr. ROS Heng**, who is being extraordinarily busy with his duties, but he still gave me warm coming time took out to hear, and guide me, even I made a mistake he still supports and keep me on the correct path and allowed me to carry out my duties. Especially, I would like to thank my lecturer **Chao Veasna**, who is being extraordinarily busy with his duties, but always support and share with me a lot of documents with many benefits.

References

[1] A. A. M. Fata and M. M. M. Aboulila, "In-building solutions using distributed antenna system based on fractal array," *Prog. Electromagn. Res. Symp.*, no. 1, pp. 984–987, 2017.

[2] N. B. Kiong and Z. A. Bin Akasah, "*Maintenance factor for precast concrete in IBS: A review*," *2011 Natl. Postgrad. Conf. - Energy Sustain. Explor. Innov. Minds, NPC 2011*, 2011.

[3] J. L. Honglin Hu, Yan Zhang, *DISTRIBUTED ANTENNA SYSTEMS.* 2007.

[4] D. Castanheira and A. Gameiro, "*Distributed antenna system capacity scaling*," *IEEE Wirel. Commun.*, vol. 17, no. 3, pp. 68–75, 2010.

[5] J. Zhang, S. Member, J. G. Andrews, and S. Member, "*Distributed Antenna Systems with Randomness*" vol. 7, no. 9, pp. 3636–3646, 2008.

[6] N. Petrovi and D. Savkovi, "*LTE Performance in a Hybrid Indoor DAS (Active vs . Passive)*" vol. 7, pp. 141–144.

[7] T. Jönsson, C. Nizman, M. Bergeron, K. Larsson, and A. Simonsson, "*Real life LTE in-building deployment demonstrating multi-cell capacity*" *IEEE Veh. Technol. Conf.*, vol. 0, pp. 0–5, 2016.

Real Time Flood Monitoring Using IoT Sensors

Ngob Duong[#1], Chea Socheat[*1]

Department of Telecoms & Networking, Faculty of Digital Engineering,
Institute of Digital Technology, Cambodia Academy of Digital Technology, Cambodia
duong.ngob@student.cadt.edu.kh
sochet.chea@cadt.edu.kh

***Abstract*— Floods are among the most common damaging natural disasters that cause significant harm to life, spread of diseases, property, and economy. Due to climate change, scientists hardly understand where the next flood calamity would be and as such cannot do any predictions to warn people. Damage to houses and buildings are prominent and cost a lot to repair. Flood monitoring is a smart way to monitor floods and prevent damage when a disaster strikes, at any moment. It is important to pre-warn residents in flood prone areas of incoming floods so that they can prepare themselves to evacuate. The implementation of a flood alert system near any flood prone area can provide critical information about the situation and can help protect properties and saves lives. In this paper, we present a real-time flood monitoring system-using ESP 8266 with Ultra Sonic sensor with firebase.**

***Keywords*— Wireless Sensor, Real-time database, Flood monitoring, Firebase, Node-Red**

I. Introduction

Due to climate change, scientists hardly understand where the next flood calamity would be and as such cannot do any predictions to warn people. Damage to houses and buildings are prominent and cost a lot to repair. Flood monitoring is a smart way to monitor floods and prevent damage when a disaster strikes, at any moment. It is important to pre-warn residents in flood prone areas of incoming floods so that they can prepare themselves to evacuate.

According to reports from the Cambodia Humanitarian Response Forum (CHF), From 01 September to 13 October 2022, flooding in Cambodia has affected an estimated 85,482 households across 74 districts in the 14 provinces of Kampong Thom, Battambang, Banteay Meanchey, Pursat, Preah Vihear, Siem Reap, Tboung Khmum, Kratie, Stung Treng, Oddar Meanchey, Kampong Chhnang, Mondul Kiri, Kampong Cham, and Kampong Speu. Fourteen people died in floods in Banteay Meanchey Province, and one in Oddar Meanchey Province.

II. Literature Review

A. ESP 12-E

ESP8266 is a low-cost and high-performance wireless SOC, which provides endless possibilities for integrating Wi-Fi functions into other systems. It can control input and output like Arduino, but the special feature is that it comes with Wi-Fi. Compared with other Wi-Fi solutions on the market, ESP is the best option for most "Internet of Things" projects! Because it is so cheap that it only costs a few dollars, it can also be integrated into advanced projects. Moreover, ESP is compatible with Arduino IDE.

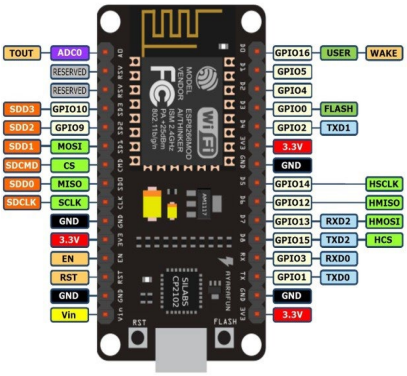


Fig. 1 ESP 12-E

ESP-12E is a member of the "ESP-XX" series. It is a miniature Wi-Fi module used to establish a wireless network connection for a microcontroller or processor. The core of ESP-12E is ESP8266EX. This module has no complicated circuits or programming so using this module is very easy.

B. Ultra Sonic sensor

An ultrasonic sensor emits a sound pulse in the ultrasonic range. This sound pulse propagates at the speed of sound through air (about 344 meters per second) until the sound pulse encounters an object. The sound pulse bounces off the object and is returned in reverse to the sensor where this "echo" is received. By measuring the time, it takes the sound pulse to travel from sensor to object and back to the sensor. The distance to the object can be calculated very accurately. This measuring principle is also called "Time of Flight", or transit time measurement.

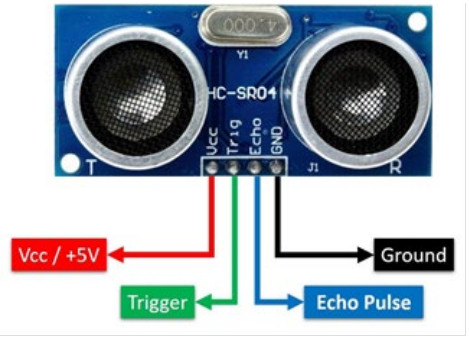


Fig. 2 Ultra Sonic Sensor

C. *Firebase*

Firebase is a toolset to "build, improve, and grow your app", and the tools it gives you cover a large portion of the services that developers would normally have to build themselves but don't really want to build because they'd rather be focusing on the app experience itself. This includes things like analytics, authentication, databases, configuration, file storage, push messaging, and the list goes on. The services are hosted in the cloud and scale with little to no effort on the part of the developer.

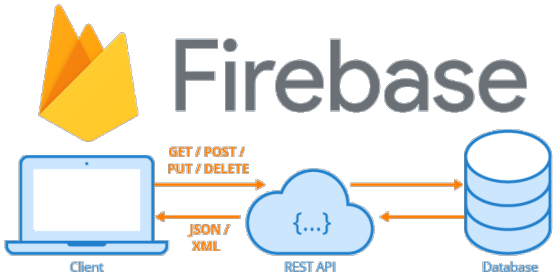


Fig. 3 Firebase function

D. *NODE-RED*

Node-RED is a flow-based programming tool, originally developed by IBM's Emerging Technology Services team and now a part of the Open JS Foundation. Node-RED consists of a Node.js based runtime that you point a web browser at to access the flow editor. Within the browser you create your application by dragging nodes from your palette into a workspace and start to wire them together. With a single click, the application is deployed back to the runtime where it is run. The palette of nodes can be easily extended by installing new nodes created by the community and the flows you create can be easily shared as JSON files.

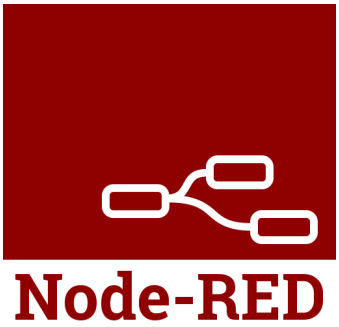


Fig. 4 Node-red (logo)

III. METHODOLOGY

A. *Block Diagram*

Fig. 5 illustrates the block diagram of Proposed System Model. All the hardware equipment and respective sensors can be addressed form the below given figure.

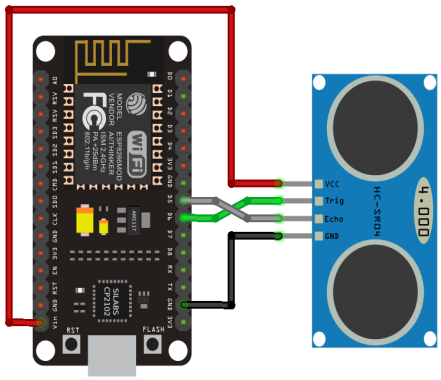


Fig. 5 Block Diagram

B. *Esp8266 with Firebase*

You can use the ESP8266 to connect and interact with your Firebase project, and you can create applications to control the ESP8266 via Firebase from anywhere in the world. In this project, we'll create a Firebase project with a real-time database, and we'll use the ESP8266 to store and read data from the database. The ESP8266 can interact with the database from anywhere in the world as long as it is connected to the internet. This means that you can have two ESP8266 boards in different networks, with one board storing data and the other board reading the most recent data.

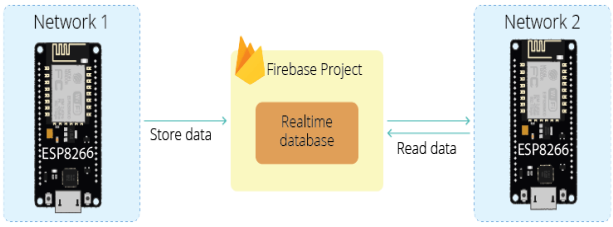


Fig. 6 Two ESP8266 boards in different networks

We also can create a web app using Firebase that will control the ESP8266 to display sensor readings or control outputs from anywhere in the world.

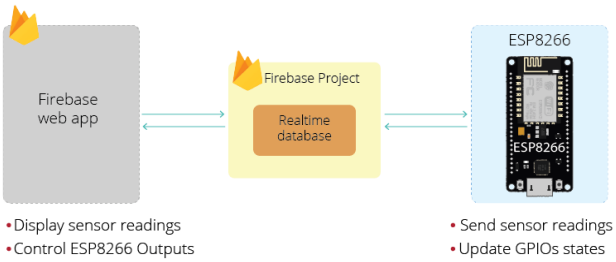


Fig. 7 web app using Firebase that will control the ESP8266

C. *Node-red with firebase*

We can provide an elegant GUI that allows you to interact with your Firebase data using Node-Red. As a side benefit, these nodes should be a natural complement to the data explorer interface offered by the Firebase Forge and a faster/easier to use tool than diving straight into the programming.

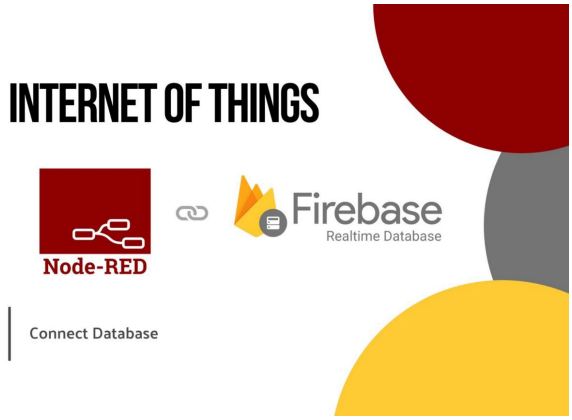


Fig. 8 Node-red with Firebase

IV. Implementation and Result

Looking at the drawbacks of the existing system, we need to introduce a new flood monitoring system which will include the sensors and hardware such as ESP 12-E and Ultra Sonic sensor. The proposed system will help people to be well prepared for flood that coming.

A. *Step of Implementation*

Step 1. Internet Connection: after power on ESP, it will start connect to internet if that it's internet connection it will move to step 2 otherwise it will try to connect to internet again

Step 2. Firebase Connection: after connect to internet ESP will connection to Real-time Firebase

Step 3. Internet Connection testing: before sensing ESP will test the internet connection again if there is not internet it will move to step 1 otherwise it will go to step 4.

Step 4. Sensing: Ultra sonic sensor will send sound to from water if the sound touches the water it will send back to ultra-sonic sensor ESP start calculate the distance between sensor and water by the time that sensor sent sound and sound comeback to sensor.

Step 5. Water level calculation: After getting distance between sensor and water, we can calculate the water level.

Step 6. Sending water level to Real-time Firebase: ESP start to send water level to Real-time Firebase, after sent data to Real-time Firebase ESP will start step 3 again to testing Real-time sensor and sending water level again.

Step 7. Monitoring System: after sending water level to Real-time Firebase we use Node-red as GUI that allows us to monitor the water level.

B. *Testing Results*

a) Water level calculated from ESP: After connected to internet ESP and Ultra Sonic sensor start calculate water level.

```
Water Level (cm): 3.55
Water Level (cm): 5.55
Water Level (cm): 3.28
Water Level (cm): 0.68
Water Level (cm): 1.54
Water Level (cm): 5.55
```

Fig. 9 Water level read from sensor

b) Sending data to Real-time Firebase: After get water level from sensor the ESP work as Micro controller and send data to firebase.

```
Water Level (cm): 19.43
PASSED
PATH: /Water level
TYPE: float
ETag: VP9zBiJamaV7h7PD24kP4GmXXmw=
------------------------------------
```

Fig. 10 send data from esp to firebase

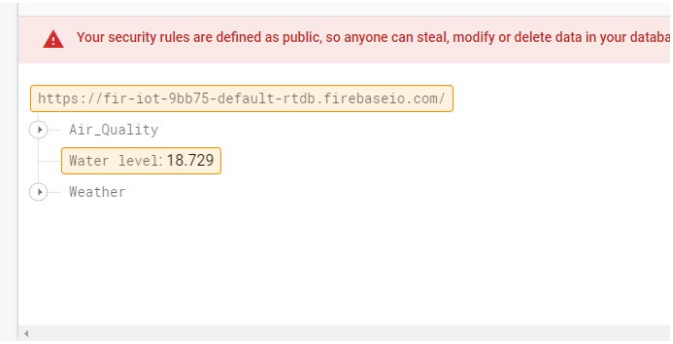


Fig. 11 Data view on real time firebase

c) *Monitoring on Node-red:* After that we use node-red for user monitoring.

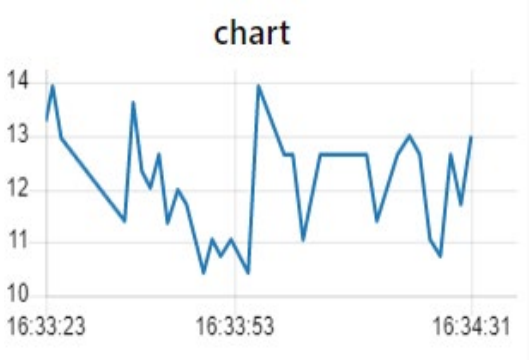


Fig. 12 Water level monitoring on node-red

V. CONCLUSIONS

Flood monitoring is a smart way to monitor floods and prevent damage when a disaster strikes, at any moment. We use ESP as a IoT broad because its power enough for this project and it's also cheap. And the future we also want to develop our project that can also use sim card internet and Wi-Fi and the same so we can access and send data even sometime that we don't have internet. In addition, we will develop our application help alert if water level come to alert level or when the water level increases too fast.

ACKNOWLEDGMENT

It is always a pleasure to remind me about those who always inspire me and gave me the possibilities to complete my internship program. Without their help from them, this whole project would not have been possible, and success on my internship. Firstly, I would like to pay full respect and gratitude to my advisor. Mr. CHEA SOCHEAT who has been trying and working hard to explain, teach, and guide me in line to finish my report on my research paper. In addition, I would like to thank my whole family who is warmly caring, supporting, and motivating me until I get successful completion of the report and all my mates who help me to check and correct this report.

REFERENCES

[1] Yupeng Huang, Wenchen Lin, Haijiang Zheng, "*A Decision Support System Based on GIS for Flood Prevention of Quanzhou City*", 2013.

[2] Yingchun Shi, Yangqi Zheng, Mao Li, "*Neural Network Model on Basin Flood Prevention Effect Assessment*", 2009.

[3] Fatimah Curumtally, Rajeev Khoodeeram, "*Real Time Flood Monitoring and Prevention Using IoT Sensors in Developing Countries*", IST-Africa Conference, 2021.

[4] HaoFang Wang, WenYan Chen, SuLin Song, "*Design of Jinan City Flood Prevention and Warning Decision-Making Support System Based on SQL Server and GIS*", First International Workshop on Database Technology, 2009.

Microwave Link Design for Long Distance Connectivity at Rural Area using Radio Mobile

Nut Kunthy[#1], Chea Socheat[*2]

Department of Telecoms and Networking, Faculty of Digital Engineering,
Institute of Digital Technology, Cambodia Academy of Digital Technology, Cambodia
[#1]kunthy.nut@student.cadt.edu.kh
[*2]socheat.chea@cadt.edu.kh

***Abstract*— Wireless communication technology with high-frequency band is used to providing high speed transmission for long distance connectivity in free space environment. This technology is applied to transmitting the information in form of electromagnetic wave is known as microwave. However, microwave is a line-of-sight communication within shortwave radio and very easily affected by atmosphere, obstacles, and other environmental conditions. The aims of this paper is to study on microwave propagation and link design for real environment at the remote area. Thus, we will create one transmission link with frequency range 5925 MHz at a distance around 13 Km to survey and apply a method for link budget calculation. Then, we will make a simulation on Radio mobile to simulate and predict communication result in real input parameters. Consequently, the link provided a good signal transmission in line-of-sight with the development of low cost and sustainable connectivity.**

***Keywords*— Line of Sight, Link budget, Microwave propagation, Fresnel Zone, Free Space Path Loss**

I. INTRODUCTION

Microwave radio refers to point-to-point fixed digital links that operate in duplex mode [12]. Microwave links generally operate between frequencies of 2 to 58GHz and it is a type of digital microwave links. Digital microwave links have many advantages such as: high tolerance against interference, high tolerance against deep fading, high signal carrying capacity ranging from 2 to 155Mbps, high frequency range (2 to 58GHz), and easy rapid installation [11].

According to another paper they developed of low-cost and sustainable wireless connectivity for long distance between the city to remote at high school or secondary school in the rural area. However, in this paper we design link for long distance connectivity at rural area by using radio mobile.

II. LITERATURE REVIEW

A. Atmospheric Effects on Propagation

1) Refracted wave: Radio links are an electromagnetic wavefront that is infinitely wide even with high-gain microwave antennas. The path that the wavefront travels is dependent on the density of the troposphere the lower portion of the atmosphere that it encounters. In a standard atmosphere, the average density decreases with altitude. The upper portion of the wavefront thus travels faster than the lower portion that is traversing the denser medium. Since the direction of propagation of an electromagnetic wavefront is always perpendicular to the plane of constant phase, the beam bends downward. This is called refraction[12].

2) Diffraction wave: Diffraction wave occurs when an impenetrable body obstructs the radio path between the transmitter and receiver. It is the important phenomena which have been studied using different approaches[12].

B. Free Space Propagation:

Radio waves are affected by the Earth and the atmosphere surrounding it. Microwave point-to-point links it is the nonionized lowest portion of the atmosphere (below roughly 20 km), called the troposphere, which is of interest. For path-planning purposes, it is useful to define a reference position where the propagation can be considered unaffected by the Earth [3]. Most RF comparisons and measurements are performed in decibels. This gives an easy and consistent method to compare the signal levels present at various points. Accordingly, it is very convenient to express the free space path loss formula, FSPL, in terms of decibels. It is easy to take the basic free space path loss equation and manipulate into a form that can be expressed in a 13 logarithmic format. For typical radio applications, it is common to find f measured in units of GHz and d in km, in which case the FSPL equation becomes:

$$\mathrm{FSPL(dB)} = 20\log(d) + 20\log(f) + 92.45 \qquad \textbf{(Eq.1)}$$

Note: d is distance of the receiver from the transmitter (km), f is frequency (GHz). On the other hand, FSPL also depends on the measure such as: d, f in meters and kilohertz, respectively, the constant becomes −87.55; d, f in meters and megahertz, respectively, the constant becomes −27.55; d, f in kilometers and megahertz, respectively, the constant becomes 32.45[12].

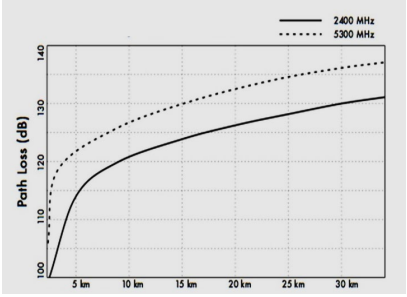


Figure 1: Free space path loss in different distance

C. *Line-of-Sight of Propagation*

Line-of-sight propagation is a characteristic of the electromagnetic radiation or acoustic wave propagation which means waves which travel in a direct path from the source to the receiver. Electromagnetic transmission includes light emissions traveling in a straight line. The rays or waves may be diffracted, refracted, reflected, or absorbed by atmosphere and obstructions with material and generally cannot travel over the horizon or behind obstacles.

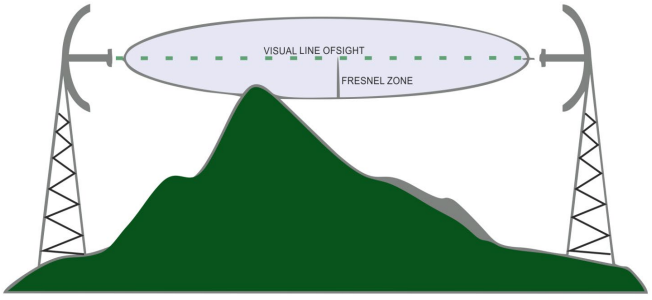


Figure 2: Line of Sight

1) Fresnel zone: Fresnel zones are used in propagation theory to calculate reflections and other losses between a transmitter and receiver. Fresnel zones are sequentially numbered and are called 'F1', 'F2', 'F3' etc. There are an infinite number of Fresnel zones, however, only the first 3 have any real effect on radio propagation. The Fresnel Zone radius is important when calculating signal loss between 2 sites. If the main signal is clear of any objects along the path (trees, hills, mountains, etc.) the path is unobstructed.

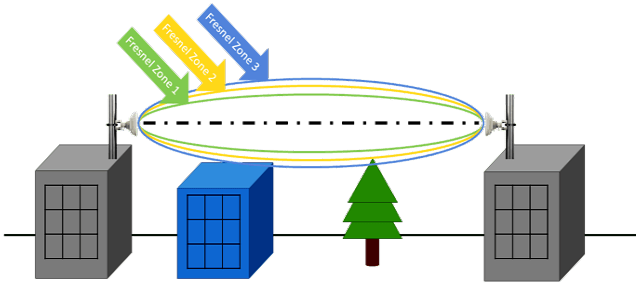


Figure 3: Fresnel Zone

2) First Fresnel Zone: an ellipsoid-shaped volume around the Line-of-Sight path between transmitter and receiver. The Fresnel Zone is important to the integrity of the RF link because it defines a volume around the LOS that must be clear of any obstacle for the maximum power to reach the receiving antenna. Objects in the Fresnel Zone as trees, hilltops and buildings can considerably attenuate the received signal, even when there is an unobstructed line between the TX and RX. Most wireless professionals work with an approach that demands that the first Fresnel zone be unobstructed, although one might be more demanding. Others demand a radius containing 60% of the total power unobstructed. The formula below calculates the first Fresnel zone [2,13].

$$r = F1 = 17.32 \times \sqrt{\frac{d1[Km] \times d2[Km]}{f \times D}} \qquad \textbf{(Eq.2)}$$

Note: d_1 is distance to obstacle from transmitter, d_2 is distance to obstacle from receiver, D for distance from transmitter to receiver (km), f = frequency (GHz) r = radius (m).

3) Fresnel Zone Clearance: The concept of Fresnel zone clearance may be used to analyse interference by obstacles near the path of a radio beam. The first zone must be kept largely free from obstructions to avoid interfering with the radio reception. However, some obstruction of the Fresnel zones can often be tolerated, as a rule of thumb the maximum obstruction allowable 60% of first Fresnel zone. If the signal path exceeds 60% clearance of F1 (First Fresnel zone), the radio signal is considered "clear line-of-sight" and will incur no diffraction loss. The radio containing 60% of the total power can be calculated[13]:

$$h_C\,(60\%) = 0.6F_1 = 10.4 \times \sqrt{\frac{d1[km] \times d2[km]}{f \times D}} \qquad \textbf{(Eq.3)}$$

Note: h_C is Line of sight clearance of first Fresnel zone (m).

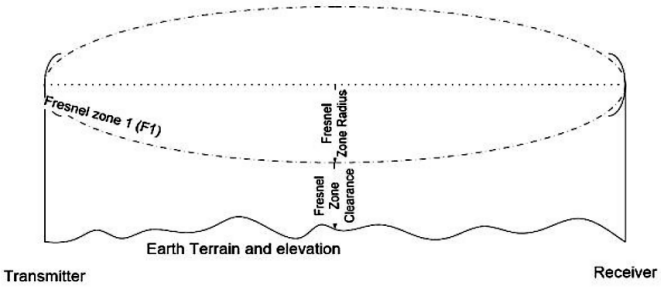


Figure 4: Fresnel Zone Clearance

4) Effective Earth Radius: It could be reasonably assumed that the direct LOS path of propagation follows a straight line; such as shown in our path profile. The reality is that the refractive properties of the earth's atmosphere can cause the path to bend slightly toward or away from the earth's surface. This curved path makes the determination of adequate obstacle clearance problematic prior to computerized propagation models, path profiles were drawn by hand on paper. To allow for a straight-line path of propagation, a correction factor, k, is applied to the earth's radius; in effect, increasing or decreasing the surface curvature to compensate for the refracted path. This corrected radius is called the effective earth radius. The k-factor used to derive the effective earth radius is a function of the atmosphere's refractive index. For a non-refractive atmosphere, in which the path of propagation is a straight line, k = 1. For the more realistic atmosphere where the path of propagation is refracted towards

the surface, k > 1. In the rare case where the path of propagation is refracted away from the surface, generally associated with a coastal sea adjacent to a large desert, k < 1. Values of k will vary by geographical region, time of year, and local weather conditions. In lieu of specific data, propagation models utilize a standard atmosphere where the k-factor is defined as: k = 1.33. The height of surface curvature at a given point along the path of propagation as a result of the effective earth radius is given by[13]:

$$h_{ER} = 0.0785 \times \frac{d1 \times d2}{K} \qquad \textbf{(Eq.4)}$$

Note: h_{ER} is height of earth radius (m).

5) Line of Sight Clearance: We can calculate the LOS clearance relative to the obstruction by[13]:

$$h_c = h_1 + (h_2 - h_1) \times \frac{d1}{D} - h_{ER} - h_o \qquad \textbf{(Eq.5)}$$

Note: d_1 is distance to obstacle from transmitter (km), d_2 is distance to obstacle from receiver (km), D is (d1+d2) distance from transmitter to receiver (km), h_1 is antenna height for transmitter (m), h_2 is antenna height for receiver (m), h_{ER} is earth radius (m), h_o is obstacle height (m) and h_c (CLOS) is line of sight clearance height (m).

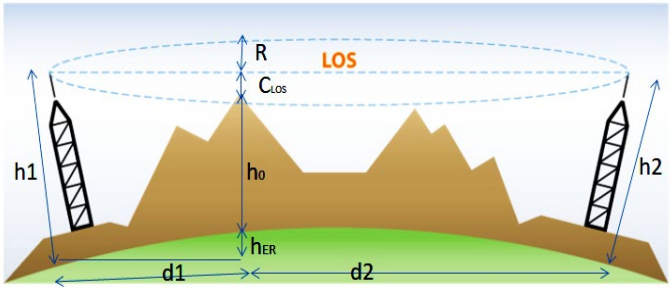


Figure 5: Path profiles in line of sight

III. MICROWAVE LINK DESIGN

Microwave link design is a methodical, systematic, and sometimes lengthy process that includes the following main activities: Loss/attenuation calculations. Fading and fade margins calculations. Frequency planning and interference calculations. Quality and availability calculations[12].

A. Link Planning

To determine the initial network topology first, the site location of the customer end sites must be determined then an initial diagram with the circuit connections and traffic capacity can be worked out. Second network diagram then initial mapwork [10,12].

1) Site Location: Really important to check and verify information on site locations. Microwave radio links allow very little inaccuracy of the site coordinates because the clearance of the beam is critical. In most cases, site coordinates need to be accurate to within a few meters [10,12].

2) Network Diagram: When the site locations have been ascertained, they should then be plotted geographically with the logical circuit connections shown to produce a network diagram. A typical GSM network showing network capacity is shown in Figure 6 [10,12].

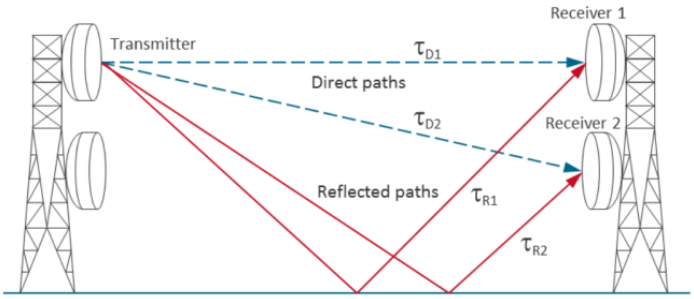


Figure 6: Typical GSM capacity diagram

3) Initial Mapwork: After determined the network connections and initial capacities, the first pass radio network diagram is required. Microwave radio links rely on there being LOS between the two ends. it should be assumed that if there is no LOS (i.e., the path is blocked by the terrain itself, some trees, or a building), the path probably will not work. to make the path of LOS work we will study on microwave propagation [10,12].

B. Link Budget Calculation

1) Power Received: A wireless link budget for a point to point radio link is the accounting of all the gain and losses from the radio transmitter (source of the radio signal), through cables, connectors and free air to the receiver. Estimating the value of the "power" in the different parts of the radio link is necessary to be able to make the best design and the most adequate choice of equipment. A simple link budget equation looks like this[14]:

$$P_r\,(dBm) = P_t\,(dBm) + G_r\,(dBi) + G_t\,(dBi) - FSPL(dB) - L_{Tx} - L_{Rx} - L_M(dB) \qquad \textbf{(Eq.6)}$$

Note: P_r is received power (dBm), P_t is transmitter output power (dBm), G_t is transmitter antenna gain (dBi), L_{Tx} is transmitter losses (dB), $FSPL$ is free space path loss (dB), L_M is miscellaneous losses, G_r is receiver antenna gain (dBi), L_{Rx} is receiver losses (dB).

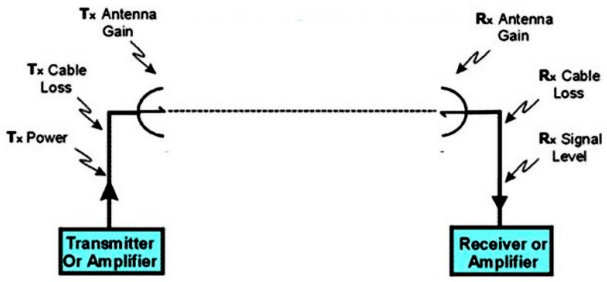


Figure 7: Parameters in radio link

C. *Effective Isotropic Radiated Power (EIRP)*

The Effective Isotropic Radiated Power (EIRP) or the Maximum Radiated Power is regulated by the national radio regulatory authority. It specifies the maximum power that is legally permitted to be send out to the free air in a specific country/area. EIRP is a measure of the effective output of a system and is expressed as equivalent to an isotropic radiating system. In simple words, this parameter tells us how strong we are allowed to send our signal in the free air. The radiated power is the result of subtracting power losses in the cable and connectors to the Transmitter Power and adding the relative "gain" of the antenna[15].

$$EIRP(dBm) = P_t\,(dBm) - Losses\ (cable,connectors)(dB) + Gt\ (dBi) \quad \textbf{(Eq.7)}$$

Calculating link budgets is all about making sure that the margin in the receiver side is higher than a certain threshold. Furthermore, the EIRP must be within regulations. The margin of a link budget can be summarized.

$$Margin = P_t\,(dBm) - Cable\ Tx\ loss(dB) + G_t\,(dBi) - FSPL(dB) + G_r\,(dBi) - Cable\ Rx\ loss(dB) - receiver\ sensitivity(dBm) \quad \textbf{(Eq.8)}$$

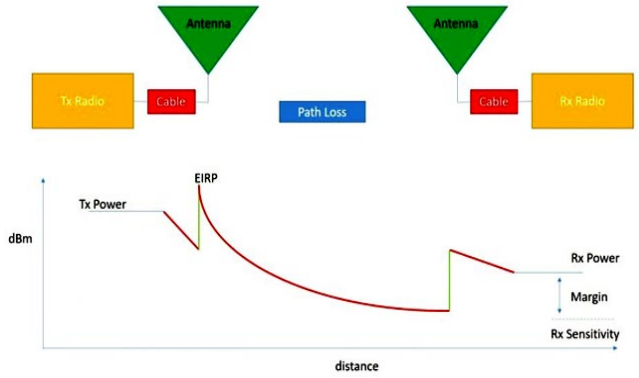


Figure 8: Link budget calculation

IV. Implementation and Result

In microwave link design, Line of Sight (LOS) is very importance factor to design the communication link in free space environment without obstacle any object or any reflection when electromagnetic wave passing through. It means to provide a good transmission performance. In this project, we were studying one link from SVR1 to SVR2 by using frequency band 6GHz which has a range 5925MHz, distance 13.36 km, transmit power 25 dBm, Antenna gain 37.3 dBi, Cable and Connector loss 0.5 dB and K= 4/3 and we saw that this link was across a factory which has height 10.2 m also as shown in the figure 9. After calculating by theoretical, we will do a simulation on Radio Mobile simulator for this link.

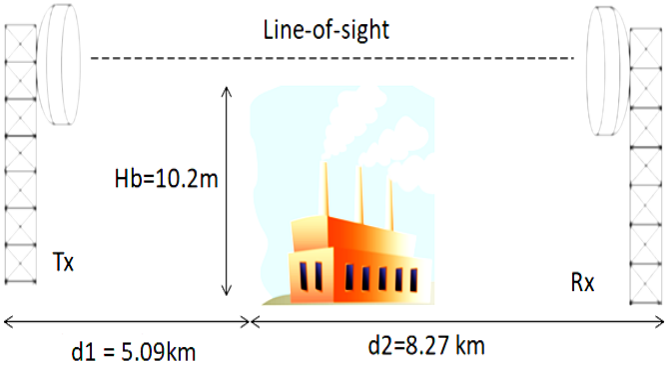


Figure 9: Topology of the project

A. *Radio Mobile Simulation*

The purpose in this point, we will learn how to use Radio Mobile, a free software that provides a detailed propagation model for radio. It allows to simulate a radio link and perform "what if?" scenarios, by changing the link parameters to predict the performance of the outdoor radio systems. Moreover, it provides us to calculate the radio link budget, predict result of radio link and analyse the simulation. We also can use this software for both Off-line (Window) and Online (Web). Radio Mobile uses the following input parameters to predict and provide a coverage map showing radio coverage. Below is some parameter that we need to know in the Radio Mobile such as: Transmitter location, Transmitter power, Frequency, Antenna Type, Antenna height, Transmitter antenna Gain, Transmission line losses, including filters and multicore, Receiver location, Receiver antenna Gain, Receiver sensitivity, Terrain and elevation data for the area. The first step need to thing when we design a link is:

1) Latitude and Longitude of Sites: Latitude and longitude is probably the most widely known and universally used method of locating a point on earth that will show the location of each our sites. We have one link from **SVR1 (Tx)** to **SVR2 (Rx)**

	Latitude	**Longitude**
○ SVR1	11.04986230	105.89967910
○ SVR2	11.04032980	106.02153210

2) *Radio Link Analysis (Link Testing):* In this section, we are going to test the radio mobile simulation to analyze the radio link in a short distance from SVR1 (Tx) to SVR2 (Rx). Importantly, this testing result will have compared to the theory result to study how are they different or the same in the form of free space environment and verifiable the radio mobile software can use to analyze and predating the radio link characteristics or not.

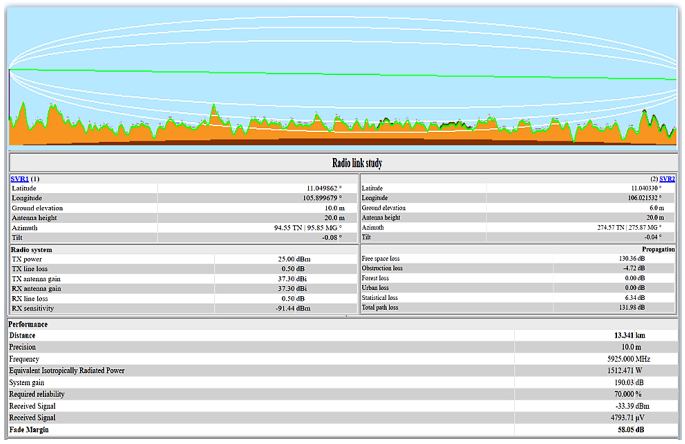


Figure 10: Radio link analysis from SVR1 to SVR2

3) Result: We will use these results to compare with the theory result to find how the radio mobile simulation are different from theory. After we simulated the radio link, we get the result above. To verify these result, we use some theories to calculate the parameter in the signal radio characteristics. By the theory we find: Rx power by the (Eq.6) we get: P_R = -32.1 dBm, EIRP by the (Eq.7) we get EIRP = 61.8 dBm or 1513.56 w, Margin by the (Eq.8) we get Margin = 59.34 dB, Free Space Path Loss (FSPL) by the (Eq.1) we get FSPL = 130.25 dB.

4) Link budget: depend on both results, we saw that it has the same result on free space path loss and EIRP but it also has the a little different on the received power and Fade Margin. Because in radio mobile simulation have the other losses like obstruction loss (-4.72 dB) and statistical loss (6.34 dB) that why it has the total path loss 131.98 dB. Thus, the Radio Mobile simulation is followed on the propagation in free space environment and we can use it to analysis and predicting the radio link. Also this software will make us to get more easy to optimize the radio link without calculation by theory.

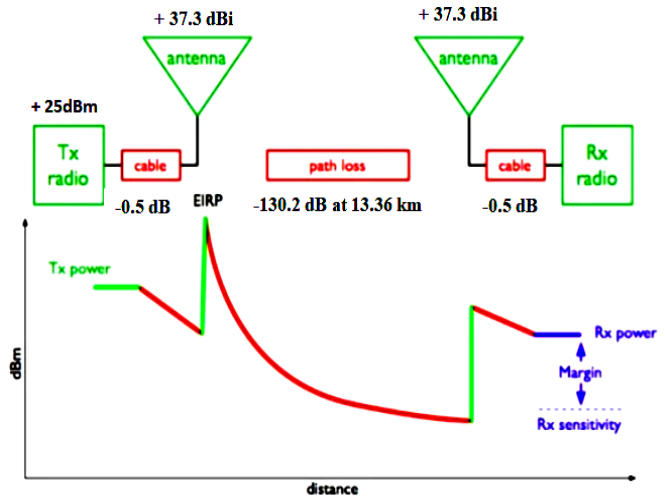


Figure 11: Link budget

V. Conclusions

This project was finished successfully by achieving our objectives, and we completed a lot of functions such as: studying microwave propagation and link design is successful, Simulation on microwave communication systems are successfully tested on Radio Mobile. Study on product specification and software to configure the system is successful. Finally, our link can apply.

Acknowledgment

I would like to express my gratitude to my advisor he is **Mr. CHEA SOCHEAT**, a Laboratory Coordinator and Instructor of school Telecoms and Networking, who is being extraordinarily busy with his duties took time to guiding me to do my research on my paper and gave me some comments and corrected along the way from the starting of writing my paper until I could complete my paper on the internship successfully.

References

[1] Kurk, Morgan, (November 17, 2021). *MICROWAVE COMMUNICATION BASICS*.CommScope experts. shorturl.at/clGM5.
[2] Alok, Kumar. (May 31, 2016). *5 Key Factors in Designing a Point to Point Microwave Link*. from Linked in. shorturl.at/nADKO
[3] Igor, Valdman. (2009). *Microwave Radio Transmission Design Guide*. ARTECH HOUSE. shorturl.at/exHOQ
[4] Huawei. (April, 30, 2015). *Product Description*. China
[5] Huawei. (2013). RTN XMC ODU. china
[6] Huawei. (2011). OptiX RTN 950 Radio Transmission System. China
[7] Huawei. (2019). IDU Quick Installation Guide for Outdoor Cabinets (APM30H&TMC11H Cabinets Ver.E). china
[8] Huawei. (2015). OptiX RTN 950A Radio Transmission System. china
[9] Pablo Angue and Juan Antonio Romo, Microwave Link of Sight Link Engineering, 2012.
[10] Harvey Lehpamer, Microwave Transmission Networks: Planning, Design, and Deployment, McGraw-Hill, 2010.
[11] Ajay R. Mishra, ADVANCED CELLULAR NETWORK PLANNING AND OPTIMISATION 2G/2.5G/3G. . . EVOLUTION TO 4G, 2007.
[12] Trevor Manning, Microwave Radio Transmission Design Guide,Artech House,2009.
[13] Campbell Scientific Companies, Line of Sight Obstruction. App. Note Code: 3RFE, 2016.
[14] T.L. Signal, Wireless Communications, Department of Electronics and Communication Engineering, 2010.
[15] Developed by Sebastian Buettrich, wire.less.dk, Edited by Alberto Escudero Pascual, IT+46, created September 2005, Radio Link Calculation Handout.

Wireless Ubiquitous Sensor Network for Home Environmental Monitoring Based on Wi-Fi Module

Sereyvath Pov[#1], Chea Socheat[*2]

Department of Telecoms and Networking, Faculty of Digital Engineering

Institute of Digital Technology, Cambodia Academy of Digital Technology, Cambodia

[1]sereyvath.pov@student.cadt.edu.kh

[2]socheat.chea@cadt.edu.kh

***Abstract*— Significant issues that often occur in a house are energy consumption and manage electronic device to maintain safety, comfortable, convenience and saving energy. In order to address this issue, a combination of experiment has been designed and developed in wireless ubiquitous sensor network for smart home environmental monitoring. The system operation process is initiated by the coordinator's sensing information reading order from the dashboard through a server with internet or directly connected and then delivering it to an end device by using radio frequency technology. The end device delivers various sensing information to the coordinator which delivers it to a server through the Internet or to a dashboard directly connected to the coordinator. As for the educational course, it is about practices such as the wireless operation process and relevant programming skills. Regarding it, the communication between coordinator and end device is designed by utilizing physical layer of RF protocol, MAC layer and network layer while the communication between server and coordinator is designed by proposing an independent protocol on TCP/IP socket and the protocol processing procedure during sensing data delivery is verified by interpretation.**

***Keywords*— Wireless Sensors, Embedded System, Wi-Fi, Ubiquitous Sensor Network, Environmental Monitoring, Radio Frequency**

I. Introduction

Recently, as interest in home network and ubiquitous has increased, low-speed, short-distance Wi-Fi communication within several tens of meters is attracting great attention. This Wi-Fi technology is considered as a pivotal technology for the implementation of Ubiquitous Sensor Network (USN). In terms of communication distance, wireless personal area network (WPAN) means only a few meters, but in the case of Wi-Fi communication distance is possible up to hundreds of meters [1]. Wi-Fi standardizes the network layer, application support layer and security and application layer as the upper layer above the physical layer and medium access control layer of the IEEE 802.11x family of standards. Due to its properties, it is mainly used for industrial and household wireless sensor networks and control of electronic devices [2]. Therefore, programming practice for understanding Wi-Fi and Wi-Fi protocol suitable for the USN environment is required.

However, there is a lack of educational equipment that can understand the Wi-Fi system and learn Wi-Fi technology. In the case of Wi-Fi equipment such as ESP32 OF Espressif and CC3200 of Texas Instruments [3], packet analysis is performed by including protocols of physical layer, MAC layer, and network layer in the operating system by using System on a chip (SoC). It is not easy to understand Wi-Fi technology because trainees can practice only simple driving programs. In this paper, we developed a software system that educates the Wi-Fi protocol flow and technology of the 2.4GHz band and implemented it on embedded hardware so that it can be used for educational purposes. That is, the protocol flow in the process of transmitting the information sensed by Ende-Device to Coordinator and its application practice were analysed. The configuration of Coordinator and Ende-Device hardware uses ESP8266-12E module as CPU [4], and ESP-NOW for communication and Node-RED for monitoring. To send the sensed information, we use sensors module [5]. In addition, MQTT server was used for messaging protocol of the server that is connected to and controlled by Coordinator [6].

II. Related Work

End-Device, which is responsible for the sensing function in the system, transmits information to the server via Coordinator [7]. The server read the End-Device data information recorded with the sensing information to translate End-Device selection process and End-Device sensing information [8]. To this end, End-Device records the information corresponding to the temperature sensor, humidity sensor, pressure sensor, tilt sensor, illuminance sensor, and IR sensor values in the data table, and the server retrieves and analyses the information recorded in the End-Device through Coordinator [9]. End-Device sensing data reading is performed in the designated memory of End-Device, and trainees can modify this data according to the purpose by referring to the existing demo program.

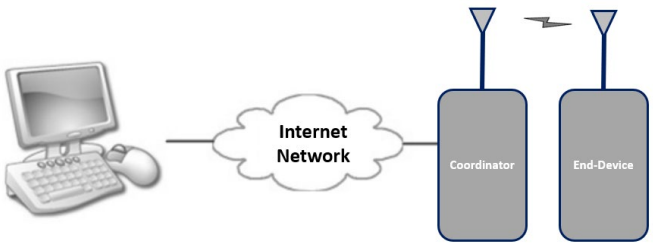


Fig. 1 Wireless Communication Network

Devices and servers can be monitored by the device itself through USB and connected to external networks through Wi-Fi. It was configured to enable this [10].

III. IMPLEMENTATION

A. Design Concept

The system implemented in this paper can be largely divided into hardware and software parts [11]. In the hardware structure, the system is largely composed of a CPU unit, a power supply unit, a monitoring unit, and an RF unit. Fig. 2 shows the End-Device of the system, and Coordinator includes an ethernet connection instead of a sensor.

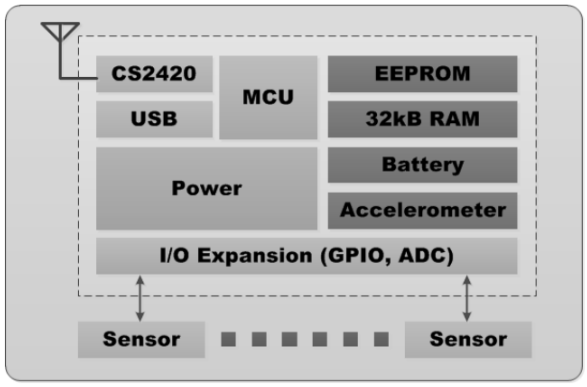


Fig. 2 End-Device's Hardware configuration

The communication part has a Wi-Fi part for performing internet communication with the server and a Node-RED part for monitoring function [12].

1) Wi-Fi part: Based on The IEEE 802.11 family of standards, it is used for Internet communication between the system and PC application programs.

2) RF protocol is used, and the frequency of the 2.4GHz band is used. Data rate is 250kbps and QPSK modulation method is used.

3) Power supply: USB VBus or battery 4.5V is used by using a constant voltage chip to down to DC3.3V.

4) CPU part: With 8bits RISC structure, instructions are simple, operation speed is fast, and it operates at 16MHz.

5) Reset: RF chip and Wi-Fi chip are reset at the same time.

6) USB interface: for system monitoring on PC Serial communication was implemented.

B. Protocol flow chart

When the firmware of End-Device and Coordinator turns on as an Ad-hoc network and becomes an element of the network, it performs a recognition process to recognize each other and a function to collect sensing information [13]. In order to read the sensing information connected to End-Device, after connecting Coordinator with Wi-Fi, the server inputs a Read command to Coordinator, and the process of reading whether the input information is normally processed, and the physical layer, MAC layer, and network in the serial monitor connected to Coordinator Function check is performed by displaying the protocol information up to the layer display values on the screen.

1) After connecting to Internet, Server start enter read command to Coordinator.

2) Upon receiving the read command, Coordinator transmits the MAC layer command to End-Device.

3) End-Device transmits MAC layer association request command to Coordinator. Then, Coordinator receives it and transmits MAC layer association response command to End-Device.

4) End-Device sends route request command of network layer and Coordinator responds with route response.

5) Coordinator sends a Read command, which is the reading of sensing information, to the payload field of the data packet of the network layer to End-Device.

6) End-Device sends a Read response, which is the result of sensing information, to the payload field of the data packet of the network layer to Coordinator.

7) End-Device or Coordinator transmits a leave packet and performs the process of leaving the network.

8) End-Device and Coordinator MAC layer Disassociation notification command is transmitted to perform the process of leaving the network.

C. Software Design

To understand the demo process, the software of the WSN monitoring system consists of End-Device, Coordinator firmware and server programs connected to the Internet.

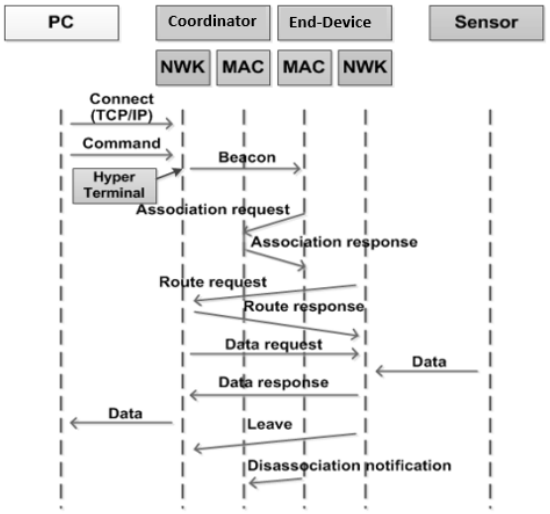


Fig. 3 Processing scenario of System

The coordinator introduced in this paper is connected to the server using wi-fi connection, and connect to the end-device through 24GHz wireless [14]. The server and coordinator comply with the middleware communication protocol, which independently defines and processes the communication protocol, and complies with Coordinator and End-Device IEEE802.11 standards. The software design is shown in Figure 4, the middleware that handles the process of

transmitting and receiving End-Device data information from Coordinator and the communication process between Coordinator and End-Device are designed.

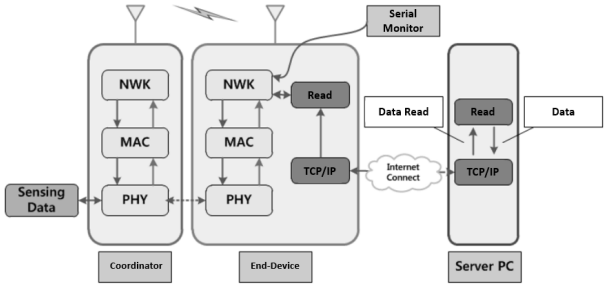


Fig. 4 The structure of the designed software

1) Server's software structure: The server introduced in this paper is connected to Coordinator through the Internet to control Coordinator. The software of the server PC was designed as a dialog base using Visual C++. This is a demo program that handles the reading function of End-Device sensing data so that trainees can practice.

2) Coordinator's Software Architecture: The read function processing received from the server performing the middleware function is received, converted into IEEE802.11 standards, and transmitted to End-Device. It performs the function of transmitting, and its configuration shown in Fig 5.

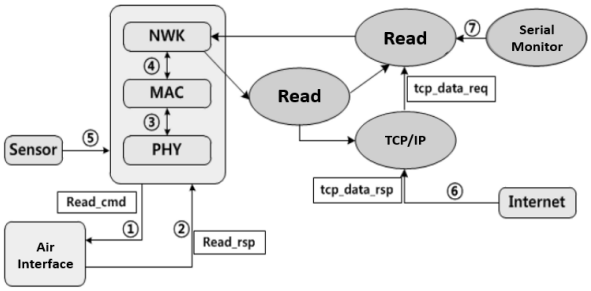


Fig. 5 Coordinator's Software structure

3) Software structure of active End-Device: End-Device receives Read command processing from Coordinator in the form of IEEE802.11 standards and transmits the sensing value. The operation process of End-Device is similar to the software structure of Coordinator, and the Internet processing part is removed and sensor processing software is added.

IV. IMPLEMENTATION

The function check of the developed system is shown in Fig. As shown in Figure 6, the linkage between the server and Coordinator, Coordinator and End-Device, and the server, Coordinator, and End-Device was verified. Similarly, the Ethernet connection between the server and Coordinator was also checked. It was checked whether Coordinator operates as a TCP Server and the server operates as a TCP Client, and Coordinator transmits information to the server for processing the read function of sensor data connected to End-Device from the server and processes in the form of IEEE802.11 standard [15]. Checked if it worked normally. The inspection process includes inspection of the server through End-Device and Coordinator and inspection of the Wi-Fi communication process between End-Device and Coordinator.

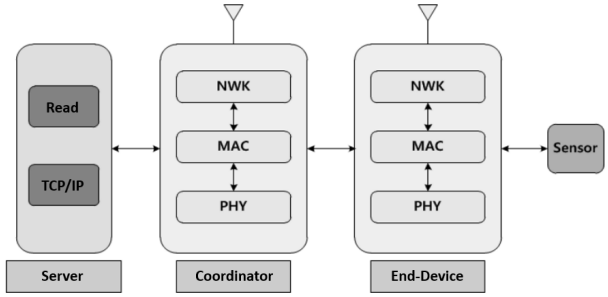


Fig. 6 Checking structure in server

A. Check on Server

The processing screen structure of the monitoring system is shown in Fig. 7 is displayed on the same screen.

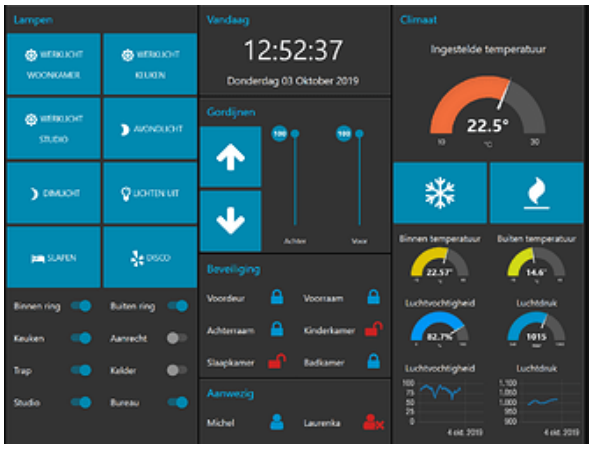


Fig. 7 Monitor processing

B. Communication between End-Device and Coordinator

By connecting serial monitor to Coordinator, there is step-by-step processing of physical layer, MAC, and network layer. As a representative example, the inspection process of the network layer is analyzed below. send read net [Address]: Read the sensing information of End-Device after data link and network layer processing [Address]: enter decimal value. Fig. 8 is the process of examining only the frame transmission process during the reading of sensing information in End-Device.

Fig. 8 Beacon frame forwarding process

Since Fig. 9 is the sensing species of the data frame of the network layer. Displays the flow and information read result.

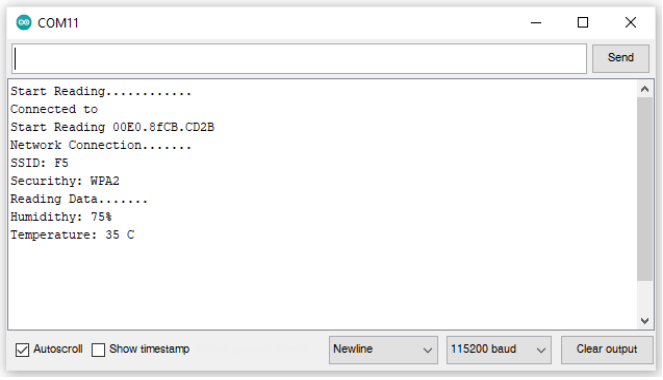


Fig. 9 The result of information reading

After the development of educational Wi-Fi system consisting of End-Device and Coordinator Understanding performance is very important. Loss between Hyper Terminal and End-Device, Coordinator When sending/receiving a kit, the primitive function is called 4 times, and one free the processing time for Mitiv is about 5ms. Also, from the sensor Packets are forwarded to End-Device and Coordinator to process MAC layer and network layer The function is called twice in End-Device and twice in Coordinator, MAC and network the time taken to process the chunk layer is about 16ms. Sensor Since the time to read the information sensed by is 3ms on average, one the time required to process the sensing information is as follows.

Time to read sensing information: 3ms

MAC, network layer processing time: 16ms * 4 = 64ms

Transmit/receive primitive processing time: 5ms *4 = 20ms

Therefore, by connecting the hyper terminal to Coordinator, the sensing type and sensor Including all the time to pass the sing value, the system's processing speed is about 90ms. Therefore, in the Wi-Fi system for education, since it takes 90ms to process my sensing data, about 11 pieces of sensing information can be processed. These results the performance of the Wi-Fi system for educational purposes is sufficiently satisfactory.

V. Conclusion

In this paper, we designed Wi-Fi software to educate the Wi-Fi protocol flow and technology in the 2.4GHz band. The designed system consists of Coordinator and End-Device nodes and can be used for educational purposes. It was implemented on embedded hardware. By developing educational equipment based on Wi-Fi communication, designing a program that can analyze protocols by layer, and technique was presented. For the configuration of Coordinator and End-Device hardware, The ESP-12E Module consists of ESP8266 SoC was used as the CPU, and radio frequency was used to send sensed information. Host and server software to control Coordinator were designed using Node-RED and MQTT Sever. In addition, as a verification of this, Coordinator was processed step by step in each layer of physical layer, MAC layer, and network layer through access to hyperterminal or server access through the Internet, and the process was analyzed. So far, only the physical layer, MAC layer, and network layer have been designed for education, but it should be designed to enable connection with commercially available Wi-Fi boards in the future.

Acknowledgment

It is always a pleasure to remind me about those who always inspire me and give me the possibilities to complete my research paper program. Without their help from them, this whole project would not have been possible, and success on my paper internship. I would like to pay full respect and gratitude to my advisor. Mr. CHEA SOCHEAT who has been trying and working hard to explain, teach, and guide me in line to finish my report on my research paper internship.

References

[1] M. A. Faraj and N. W. Boskany, “Intelligent Traffic Congestion Control System using Machine Learning and Wireless Network”, *UHD Journal of Science and Technology*, vol. 4, no. 2, pp. 123-131, 2020.

[2] Baghaei, Nilufar and Hunt, Ray, “IEEE 802.11 Wireless LAN Security Performance Using Multiple Clients,” Proceedings of the 12th IEEE International Conference on Networks, 2004.

[3] IDA HÜBSCHMANN 2020, *nabto,* accessed 14 September 2020

[4] S. Sen, J. Koo and S. Bagchi, "*TRIFECTA: Security Energy Efficiency and Communication Capacity Comparison for Wireless IoT Devices*", *IEEE Internet Computing*, pp. 74, 2018.

[5] M. Khan, A. Khan, K. A. Khan and K. M. Khan, "*Comparison Among Short Range Wireless Networks: Bluetooth Zigbee & Wi-Fi*", *Advances in Computer Science and Engineering*, no. 4, pp. 19-28, 2016.

[6] S. Neupane, *A Comparative study of Wireless Star Networks Implemented with Current Wireless Protocols*, 2019.

[7] T. N. Hoang, S.-T. Van and B.D. Nguyen, "*ESP-NOW Based Decentralized Low-Cost Voice Communication Systems for Buildings*", *International Symposium on Electrical and Electronics Engineering (ISEE),* 2019.

[8] A. Maier, A. Sharp and Y. Vagapov, "*Comparative analysis and practical implementation of the ESP32 microcontroller module for the Internet of Things*", *Internet Technologies and Applications (ITA)*, pp. 143-148, 2017.

[9] A. Škraba, A. Koložvari, D. Kofjač, E. Semenkin, V. S. Anovov and R. Stojanovic, "*Prototype of Group Heart Rate Monitoring with ESP32 Comparison to ESP8266*", *Mediterranean Conference on Embedded Computing*, 2019.

[10] S. Suherman, "*WiFi-Friendly Building to Enable WiFi Signal Indoor*", BEEI, vol. 7, no. 2, 2018.

[11] N. Kolban, Kolban's Book on ESP32, [online] Available: https://leanpub.com/kolban-ESP32.

[12] T. Herawati and F. F. T. P. W Dan Imansyah, "*Analysis Performance Wi-Fi di Fakultas Pertanian Universitas Tanjungpura menggunakan Aplikasi G-NET WiFi*", *Jurnal Teknik Elektro Universitas Tanjungpura*, vol. 2, no. 1, 2020.

[13] T. Istanto and F. X. Manggau, "*Analysis Of The Power Of Wifi Signals On The Informatics Engineering Laboratory Of Musamus University Using Insidder*", *International Journal of Mechanical Engineering and Technology (IJMET),* vol. 9, no. 13, pp. 266-272, 2018.

[14] M. Collotta, P. G. Collotta, T. T. Collotta and T. O. Collotta, "*Bluetooth 5: A Concrete Step Forward toward the IoT*", *IEEE Communications Magazine*, vol. 56, 2018.

[15] M. Ahsan Nur-A-Alam, M. Based, J. Haider and E. Rodrigues, "*Smart Monitoring and Controlling of Appliances Using LoRa Based IoT System*", Designs, vol. 5, no. 17, 2021.

Expanding Cellular Network to Improve Signal Coverage in Remote Areas

CHEA Touchvannchannvutha[#1], CHEA Socheat [*2]

Department of Telecoms & Networking, Faculty of Digital Engineering,
Institute of Digital Technology, Cambodia Academy of Digital Technology, Cambodia
[1]touchvanchannvutha.chea@student.cadt.edu.kh
[2]socheat.chea@cadt.edu.kh

***Abstract*— Due to the increasing population of user equipment (UE) and limitation of base stations, the capacities of mobile service and signal coverage is not enough for user. Moreover, there are a lot of places in the country which the cellular service cannot coverage. Meanwhile, 5G technology should be operated in the future, as we have known that 5G signal cannot provide coverage more than 1Km range so the need of 5G base stations is a big matter in Cambodia. To reach out the need in our country in this field, we have to build more sites to support mobile operating services for UE with good quality and a good key performance indicator (KPI). After installing the base stations, the KPI testing results show that we received better performance for UE outside/inside the building and the capacities of operating service are more widely, together with the wide coverage service to customers.**

***Keywords*— Cellular Network, Radio Access Network, Mobile communication, Long Term Evolution, Telecommunication.**

I. INTRODUCTION

Telecommunication networks are the transmission systems that transmit information in both analogy or digital form between various sites using electromagnetic or optical signals. The information may consist of audio or video data or other type of data. It is also a good solution for a long distance communication via an electrical system. As technology advances from 1G to 5G, it makes easier for us to connect with the rest of the world. 2G and 3G technologies provide only on voice calls and SMS, but cannot send big data or video, easy to interfere from outside and low speed as well. 4G technology is now widely available and offers a high speed that approximately ten times faster than 3G. To on-air each site, we have many teams to contribute such as planning, installation, optimization, drive testing, and walk testing [1].

According to the increasing a huge number of UE requirement data usage and a big size of data transmission, the mobile operators cannot provide on demanding request of the capacity and services because of the limitation of the data capacity on the stations. In other cases, some parts at the remote areas in Cambodia don't even have 2G or 3G signal to use. Likewise, in rural areas and/or city, mobile operators also cannot provide a good signal coverage and services for a big amount of population. Thus, a better solution to solve these problems is to build the new base stations, including BTS, NodeB, eNodeB, and gNodeB to increase the mobile service in Cambodia.

II. RELATED WORK

There are up to 4 types of macro sites that Cambodia needs, such as BTS, Node, eNodeB, and gNodeB. A base transceiver station (BTS) is a fixed radio transceiver in any mobile network. The BTS links mobile devices to the network (2G). It transmits and receives radio signals from mobile devices, converts them into digital signals, and then routes them through the network to other network terminals or the Internet. Not too different from BTS, radio base stations recognized as NodeB, Evolved NodeB (eNodeB), and Next Generation Node B (gNodeB) enable mobile devices to connect to 3G, 4G, and 5G mobile networks [2]. The 3G (Universal Mobile Telecommunication System) network, the 4G LTE (Long Term Evolution) network, and the 5G NR (New Radio) network are all home to Node B. In 5G networks, there are two different types of radio base station nodes: gNodeB and ng-eNodeB. The mobile device can connect to either of these nodes to access the 5G core network. While eNodeB enables connections between 4G LTE devices and the 5G core network using the 4G radio interface, gNodeB enables connections between 5G phones (devices) and the 5G core network using the 5G radio interface [2].

All types of sites can be on the same site, but the operator made them into different categories like BTS for 2G, NodeB for 3G, eNodeB for 4G, and gNodeB for 5G. There is another reason for constructing a new site or colonization site. A site that can provide coverage for different services for UE but not all operator customers, and another reason is that with the increasing population and the arrival of 5G technology, we need more sites to support it because logically, 5G can provide coverage ranging from 500 to 1000 meters, and as a result, Cambodia does not have enough coverage sites.

III. METHODOLOGY

A. Main Concept

In Fig.1 depicted the process milestone to build a new macro site. So, once they have the necessary information and TRC approval, all sites can begin construction. First, the operator sends a request to the telecom construction subcontractor or company, who then works on it. Sub-constructor will send their site engineer to conduct a site survey of the specific location to build the macro site after receiving a site request from the operators. Simultaneously, subcontractors must design and build antenna tower for their request. After they obtained the building site, TRC and BOQ approval, they

can begin construction on the tower. At the same time, modifying the power source is important, so plan ahead of time. After the tower has been completed and well modified by the operator, telecommunication devices such as antennas and RRUs will be installed.

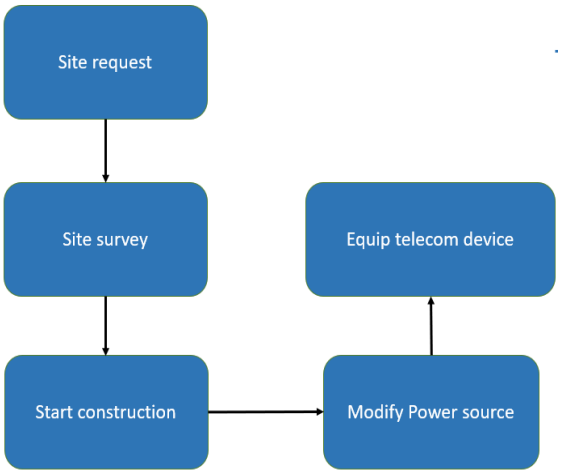


Fig. 1 Project milestone of building new sites

In general, BTS, NoteB, and eNodeB all work in the same flow and are separated in power BTS and transmission flow. Fig. 2 represents the working flow of the devices on the site; the operation of telecom technology is dependent on the power of electricity (AC), but all telecom devices are powered by DC, so we must convert AC to DC using rectifiers.

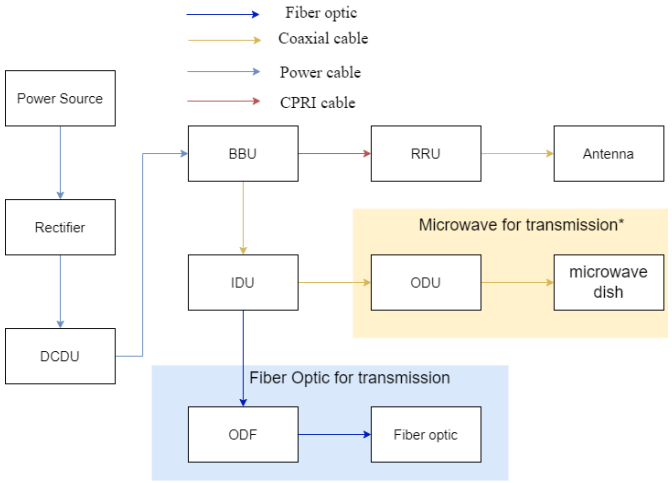


Fig. 2 Diagram flow of the site

As a result, it travels from the power source to the ACDB (Alternating Current Distribution Box), then through a rectifier to convert to DC and the DCDU (Direct Current Distribution Unit). In this step, there are some devices to include, like batteries or rectifier modules. It will go to BBU after it has converted the electricity, but BBU is connected to transmission IDU (The Indoor-Unit of the microwave system) for transmission. Thus, Data from the IDU is transformed by the ODU into an RF signal for transmission. Additionally, it transforms the distant end's RF signal into appropriate data for transmission to the IDU. ODUs are weatherproofed units located on top of a tower and can be either directly or indirectly connected to a microwave antenna by way of a waveguide. Microwave ODUs are typically made for full-duplex operation with independent transmit and receive signals. This corresponds to a "pair" of frequencies on the airside interface, one for broadcast and the other for receive. "FDD" is the name for this (Frequency Division Duplexing). Finally, it connects to the microwave dish. On the other hand, BBU normally will be connected straight to RRU by optical fiber per product specifications (based on the technology the operator wants to configure, like 2G, 3G, or 4G), and RRU will be connected to the antenna for coverage by coaxial cable.

IV. Implementation

This session will cover the requirements of the main point, technologies, equipment, and software that have used to complete this project after discussed on issue and project goal. The operator will devise a solution based on their issue. There are two ways to address that: either by building a new site or by collocating. The collocation site is operating devices that installed and coverage on the old operator site. By the way, building a new site is creating a new site.

A. Power System

This project's main component is electricity. As previously stated, all telecom products could operate on a -48V DC. The site could not function without electricity, but not all types of electricity can run on the site. Because the AC supplied by EDC in Cambodia was 220 volts, it could never be used at the site; therefore, we have to do something to convert AC to DC, which uses -48 volts. Anyways, there is seldom EDC electricity in the remote area, so in a remote area, we separate the power system into two parts such as Diesel Generator (DG) and solar. The site has to set up a solar system with a large number of solar panels to supply electricity during the daytime and DG was used at night time. But we have to alert and remind the O&M team on that DG too because DG also really needs maintenance in every 300 hours to 500 hours DG generator has to do the oil change. Fig. 3 and Fig. 4 are the power system set on the remote site in Cambodia.

Fig 3: Solar panels for power system

Fig. 4 Diesel Generator (DG)

The solution was to add rectifiers (Fig. 5) to convert AC power from the power system into DC power and provide a DC source to the UPS by using ACPDB or ACDB. An uninterruptible power supply (UPS) or battery is used to protect hardware, telecommunications equipment, or other electrical equipment where an unexpected power outage could result in injuries or data loss [5]. Devices that experience power issues like surges and power sags can be protected by both UPS and battery backups. Both choices offer protection from damage to the internal components, operating system corruption, and corruption of unsaved data. But there is a significant distinction between UPS and battery backups. A battery backup is not always activated by the process of filtering out the power of brownouts, flashing lights, and power surges. However, a UPS will filter that electricity and guarantee a steady supply of power to critical equipment that must continue operating and processing. otherwise Batteries discharge as DC as well, however, the UPS converts AC to DC for charging, whereas appliances require AC. The UPS will still convert DC to AC if the electricity goes out. The rectifier connects to batteries or UPS for backup in case there's no electricity in the site, which is used to back up the power into the battery [4]. DC power is typically supplied to equipment at 48 VDC. The rectifier connected to the UPS (for charging) is an electrical device that converts AC power to DC power, which flows in only one direction. The DC power is then supplied to the telecom equipment's batteries, which protect it when the AC power fails. Generally, most of the sites may use EDC power to support devices on the site. In any case, because EDC power cannot reach the site location in remote areas, the operator decides to use solar and DG instant.

Fig. 5 Rectifier

B. Signal Source

Logically, the source of macro sites does not take the signal link or service from others; it creates the signal and coverage service by itself. BBU is the heart of the node and made the node work because BBU provided S1 between E-UTRAN and NodeB/eNodeB and X2 port to communicate between nodes; especially, the service of the site that has to have coverage was decided by BBU; it provides the CPRI interface via an optical link from RRU. BBU also gives access to clock ports for synchronizing time, alarm monitoring ports for keeping an eye on the environment, and a Universal Serial Bus (USB) port for commissioning with a USB flash drive [2].

By the way, even though BBU is the heart of the node; RRU and RRH generated the signal. The step flow of the site is analogous to AC electricity coming from the site's power system, passing through a rectifier, which converts it to DC power, and then continuing to BBU to ensure signal integrity and manage the entire site, such as connecting with IDU for transmission; the other mission is to provide communication to the core network via S1 interface (NB/eNodeB to MME) and X2 interface (eNodeB to eNodeB), thus it connects to RRU to create a signal link and pass the signal frequency to the antenna for coverage signal [2]. The baseband unit (BBU) and RRU are linked together by an optical cable. This is since RF coaxial cable's signal loss over distance would become significant when RRU is frequently deployed far from BBU. Compared to RF coaxial cable, optical cable has significantly less loss and is more affordable. This link employs CPRI (Common Protocol Radio Interface). Multiple RRUs can be supported by a single BBU. Depending on the BBU's capability and the deployment requirements, the precise number varies [2].

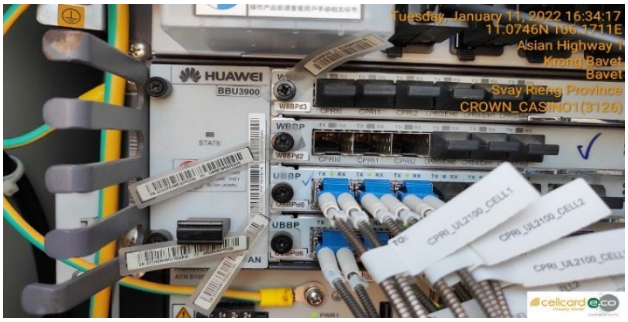


Fig. 6: Huawei BBU3900

Fig. 6 show that the site is using BBU cards (UBBP) that are connected to RRU by CPRI for coverage 4G and at the same site they also use WBBP cards too for coverage 3G in that site. Therefore, that means that that site can be represented as NodeB and eNodeB for coverage in that area. All cards that using on the site are not the same based on product specification of services and there are many vendors to supply telecoms devices to the operator so the cards that are used on the site are also different. For example, as in this implementation, the site is using BBU Huawei device so that's why they using WBBP for 3G, and UBBP (for 4G); but there are also other vendors like ZTE, NOKIA, Alcatel, etc. And the cards that are used in that are also different, too.

Building a node is not all just installing telecoms devices and that's all. Before we start installing all of those, we have to do other field things such as construction and designing towers and cabinets. To prevent dangers occur we have to install grounds, arrestors, etc. In additional installation besides telecoms devices based on Fig. 7 such as (1) Checking the installation location for the antenna and feeder system. (2) Familiarize yourself with the engineering design documentation. (3) Choose an installation strategy Lift the antenna after installing the sealing window. (4) Put the TTA in place, cut the feeders, and affix the temporary labels. (5) Prepare feeder connectors and affix labels to the feeders' outdoor portion. (6) Connect the lightning arrester to the feeders and (7) install jumpers inside Create and affix labels [3]. And the installation is also based on the overall architecture Fig. 8. As we have known that sites have some parts are the same excluding BTS, NB, and eNodeB the difference is just on the configuration of technology and the hardware specification of the device based on technology.

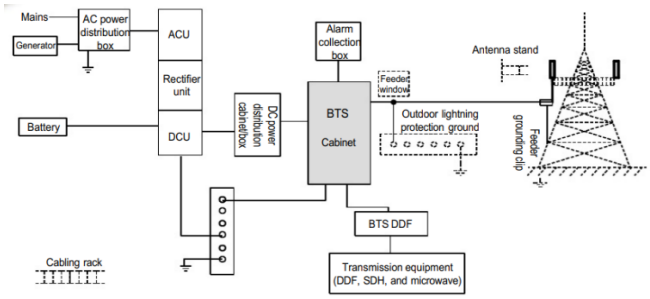


Fig. 7 Earth of the site

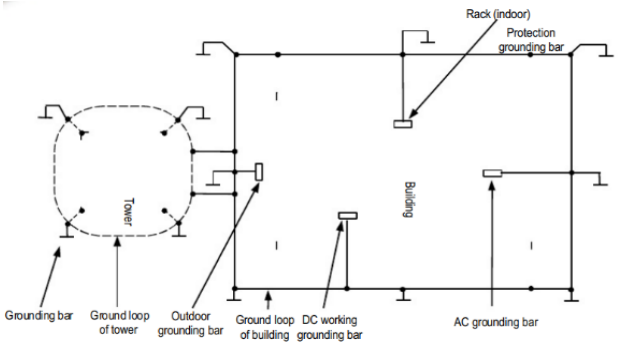


Fig. 8 Overall system architecture of the site

V. Conclusions

After the installation was fully completed, the service also cannot coverage yet, but the project is on airing. It means after all devices were equipped, this project will pass to optimize team and walk-test team to do testing and other configuration and then pass to the core network team, monitoring, and O&M team to process the project. Building a new site is hard, but it is very useful to expend mobile signal coverage and services to meet the requirement from UE. Installing the collocation site is also a good solution by separating tenants for nowadays, but will be worst in the future.

Acknowledgment

It is always a pleasure to remind me for those who always inspire and give me the possibilities to complete my research paper. This project would not possible and success without their help. Firstly, I would like to express my gratitude to my advisor, Mr. CHEA SOCHEAT, a laboratory coordinator for taking the time to advise and provide me suggestions along the way from the beginning until it is successfully completed. Last but not least, I also want to express my sincere thanks to Edotco for continually inspiring and guiding me while allowing me to study, learn, and explore this topic.

References

[1] Kerstin Peterson, "*Business Telecom Systems: A Guide to Choosing the Best Technologies and Services*", CRC Press, 2017.

[2] "LTE S1 Interface: LTE RAN to Evolved Packet Core." - CableFree, www.cablefree.net/wirelesstechnology/4glte.

[3] Huawei, "*Hardware Installation of BTS*", Engineering Center Service Management Dept, 2004.

[4] Huawei, "*Commissioning Guide*", Huawei Technologies, 2020.

[5] Alexander Kukushkin, "*Introduction to Mobile Network Engineering: GSM, 3G-WCDMA, LTE and the Road to 5G*", WILEY, 2018.

Distributed Antenna System In-Building Solution To Improve LTE signal Coverage

Voeun Soklin[#1], Chea Socheat[*2]

Department of Telecoms & Networking, Faculty of Digital Engineering,
Institute of Digital Technology, CADT, Cambodia
[#1]soklin.voeun@student.cadt.edu.kh
[*2]socheat.chea@cadt.edu.kh

***Abstract*— In-building solutions (IBS) is a telecommunications solution that is used to extend and distribute the cellular signal of mobile operators within a building to provide a good quality LTE network for indoor environments. In general, most of the new buildings can't receive a good signal coverage to be used for their requirement. In this paper, we will present a Distributed Antenna System (DAS) that can be used to efficiently increase the capacity of LTE network coverage in the building. After we installed the Distributed Antenna System In-Building Solution we received a better LTE service in the whole building that can be got a high-capacity LTE network and strengthen the quality of mobile signals inside the building, also providing smooth and efficient wireless communication for users.**

***Keywords*— Telecommunication, In-building solutions, Long Term Evolution, Distributed Antenna System, mobile signals.**

I. Introduction

In-building solutions (IBS) is a telecommunications solution which is used to extend and distribute the cellular signal of mobile operators within a building with high-quality mobile communication for indoor environments such as offices, shopping malls, hospitals, and airports; where the coverage, capacity, or quality would not have been satisfactory. Problems inside Buildings have too many types such as high Call drops when the building has above the 4th or 5th floors (Due to Multi-cell Hand over) also high Bit Error Rates due to Multipath propagation, Water refraction, Interference from other cell sites of the same operator or other operators one more thing it will no network Coverage - Basements, Ground Floors, etc. (Penetration loss) and subscriber base increases, if the deployment of new BTS sites is not possible.

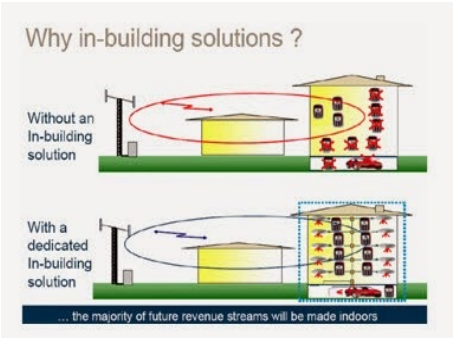


Fig. 1 Why do we need IBS?

II. Literature Review

A. *Long Term Evolution (LTE)*

LTE is a 4G standard developed in 2009 and deployed in 2010. It was considered as 3G standard by 3GPP back in 1998 but was dropped in favour of wideband CDMA because LTE requires a large amount of baseband processing, which was not commercially viable in 1998. It supports such as the speech service using Voice over IP. Another view also can support calls. It brings new features and enhancements, such as carrier aggregation, enhanced downlink control channel, advanced MIMO technique, and more. Its performance can reach download rates of up to 299.6Mbit/s and upload rates of up to 75.4 Mbit/s, and RAN latency is lower than 5ms for small IP packets. Long-term evolution (LTE) is selected as the next-generation broadband wireless technology for 3GPP and 3GPP2. The LTE standard supports both FDD (frequency division duplex), where the uplink and downlink channel are separated in frequency, and TDD (time division duplex), where uplink and downlink share the same frequency channel but are separated in time. After Rel 8, Rel 9 was a relatively small update on top of Rel 8, and Rel 10 provided a major step in terms of data rates and capacity with carrier aggregation, higher-order Multi-Input-Multi-Output (MIMO) up to eight antennas in downlink and four antennas in an uplink.

B. LTE Architecture

LTE is predominantly associated with the radio access network (RAN). The eNodeB (eNB) is the component within the LTE RAN network. LTE RAN provides the physical radio link between the user equipment (UE) and the evolved packet core network. The system architecture evolution (SAE) specifications define a new core network, which is termed as evolved packet core (EPC) including all internet protocol (IP) networking architectures Evolved NodeB (eNB): Provides the LTE air interface to the UEs, the eNB terminates the user plane (PDCP/RLC/MAC/L1) and control plane (RRC) protocols. Among other things, it performs radio resource management and intra-LTE mobility for the evolved access system. At the S1 interface toward the EPC, the eNB terminates the control plane (S1AP) and the user plane (GTP-U).
Mobility Management Entity (MME): A control plane node responsible for idle mode UE tracking and paging procedures. The Non-Access Stratum (NAS) signalling terminates at the

MME. Its main function is to manage mobility, UE identities, and security parameters. The MME is involved in the EPS bearer activation, modification, and deactivation process, and is responsible for choosing the SGW for a UE at the initial attack and at the time of intra- LTE handover involving core network node relocation. PDN GW selection is also performed by the MME. It is responsible for authenticating the user by interacting with the home subscription server (HSS). Serving Gateway (SGW): This node routes and forwards the IP packets, while also acting as the mobility anchor for the user plane flow during inter-eNB handovers and other 3GPP technologies (2G/3G systems using S4). For idle state UEs, the SGW terminates the DL data path and triggers paging when DL data arrives for the UE. Packet Data Network Gateway (PDN GW): Provides connectivity to the UE to external packet data networks by being the point of exit and entry of traffic for the UEs. The PDN GW performs among other policy enforcement, packet filtering for each user and IP address allocation. Policy and Charging Rules Function (PCRF): The PCRF supports policy control decisions and flow-based charging control functionalities. Policy control is the process whereby the PCRF indicates to the PCEF (in PDN GW) how to control the EPS bearer. A policy in this context is the information that is going to be installed in the PCEF to allow the enforcement of the required services. Home Subscriber Server (HSS): The HSS is the master database that contains LTE user information and hosts the database of the LTE users [08].

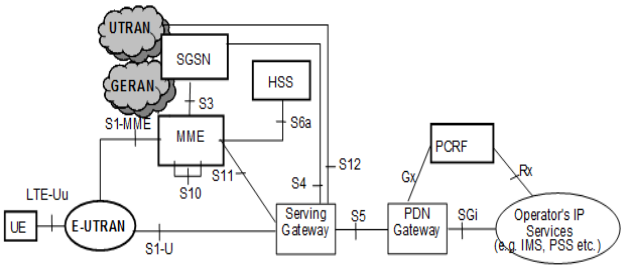


Fig. 2 LTE Architecture

C. In-Building Solution (IBS)

In-Building Solutions (IBS) are telecom network solutions comprising a distributed antenna system (DAS) enhancing coverage and capacity inside the buildings with weak or no telecom signal. It also increases telecom operators' revenue, Quality of Experience, and customer retention. There are several technologies in the market that could be utilized to provide In-Building Coverage Enhancement [06]:

1) The site survey is the process of collecting all data about the target area by visiting directly and collecting data such as geography visual inspection of the facility and prediction of the future changing environment of the site, especially the coverage area by using drive testing at the target area to see whether there is any interference or not, check RF propagation, Determine locations for equipment, confirm antenna locations, Measure interference thresholds, confirm cable routes, confirm cable lengths, check power, security and Site Photos.

2) Planning is the core part of the whole process in this project because it involves many people and if there is any problem in planning it will affect another part of Site selection, Antenna planning, Coverage planning, Capacity planning (cell planning), Link budget planning for the installation of IBS, we need to install the antenna placement (antenna estimation and antenna location), wiring diagram, and location of DAS).

3) Optimization is performed to improve the performance of the network with existing resources. The goal is to better utilize existing network resources, solve existing and potential problems, and identify possible solutions for future planning. Through Radio Network Optimization, the service quality and resource usage of the network are greatly improved to achieve a balance between coverage, capacity, and quality.

III. Methodology

A. In-Building Solution Walk Testing

Walk testing is a tedious and time-consuming task for mobile network operators, typically required to verify signal levels and the quality of service provided by a newly deployed cell site. Two traffic models were tested: A Quality of Service (QoS) based model and the typical Full Buffer model.

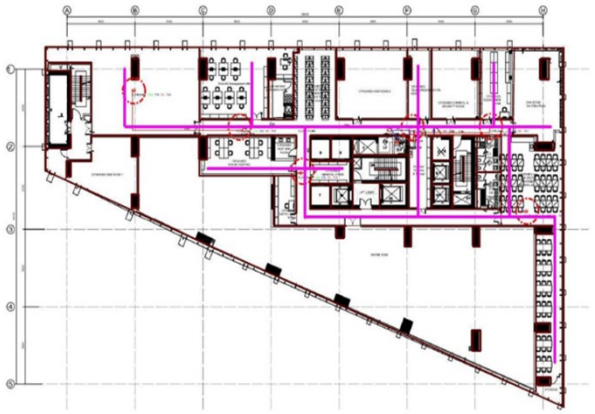


Fig. 3 IBS walk testing mapping

B. KPI Performance

In this analysis testing project, we need the most important things to complete such as: For analysis of the performance in mobile networks, we use software tools such as Genex assistant to identify the problem and analyse signal coverage. Main parameters that we use in analysis such PCI, RSRP, RSRQ, SINR, and Throughput which help us to track networks accurately in specific areas [04].

1) ***Physical Cell Index (PCI)*** determines the cell ID group and cell ID sectors. There are 168 possible cell ID groups and 3 possible cell ID sectors, so 3*168=504 possible PCI. When cell ID is set to auto, the demodulator will automatically

detect cell ID. When cell ID is set to manual, the PHY-layer cell ID must be specified for successful demodulation.

2) ***Reference Signal Received Power (RSRP)*** is the most fundamental measurement, this is basically a power measurement for a single subcarrier, so the value doesn't change with bandwidth or the number of RBs currently assigned for PDSCH. It is the power of the LTE Reference Signals spread over the full bandwidth and narrowband. RSRP gives you an idea of the strength of the signal it gets from the network, but it is not a clear indication of how good the signal quality is. RSRP=linear average power of reference signal and its range is around -44dbm (good) to -140dbm (bad).

TABLE 1
RSRP SIGNAL LEVEL AND LEGEND

Legend	Range (dBm)	Service	Count	ServRSRP
	<-100	Cell Edge	2623	5.05%
	-90 to -100	Mid Cell	15844	30.51%
	-80 to -90	Good	20946	40.33%
	>=-80	Excellent	12526	24.12%

3) ***Reference signal received quality (RSRQ)*** is always a negative value in dB. The higher RSRQ is, the better signal quality is if bandwidth and number RB allocation are the same. RSRQ depends on bandwidth and the number of RB. The RSRQ measurement provides additional information when RSRP is not sufficient to make a reliable handover or cell reselection decision. To indicate the quality of the received signal and its range is typically -19 dB (bad) to -3 dB (good).

TABLE 2
RSRQ SIGNAL LEVEL AND LEGEND

Legend	Range (dB)	Service	Count	ServRSRQ
	<-20	Cell Edge	168	0.32%
	-15 to -20	Mid Cell	4030	7.74%
	-10 to -15	Good	30225	58.08%
	>=-10	Excellent	17619	33.86%

4) ***Throughput:*** If we know the throughput and bandwidth levels for our network, we have valuable information for the assessing networking performance. Throughput tells us how much data was transferred from a source at any given time and bandwidth tells us how much data could theoretically be transferred from a source at any given time. Speed is one of the most important things used to measure network performance and we use throughput and bandwidth to measure it.

TABLE 3
RSRP SIGNAL LEVEL AND LEGEND

Legend	Range (Mbps)	Service	Count	ThroughputDL
	<0M	Bad	0	0.00%
	0M to 10M	Acceptable	29972	70.02%
	10M to 20M	Midcell	5845	13.66%
	20M to 50M	Good	5448	12.73%
	50M to 90M	Very Good	1438	3.36%
	>=90M	Excellent	101	0.24%

IV. IMPLEMENTATION AND RESULTS

A. Step of Implementation

On a rollout or optimization project where radio network optimization consultants are focusing on the problems and troubleshooting activities, this service is to complement optimization consultants in collecting all relevant and necessary information from the network by performing field measurements. Activities included are Site survey, Planning, Installation, Testing, and Optimization.

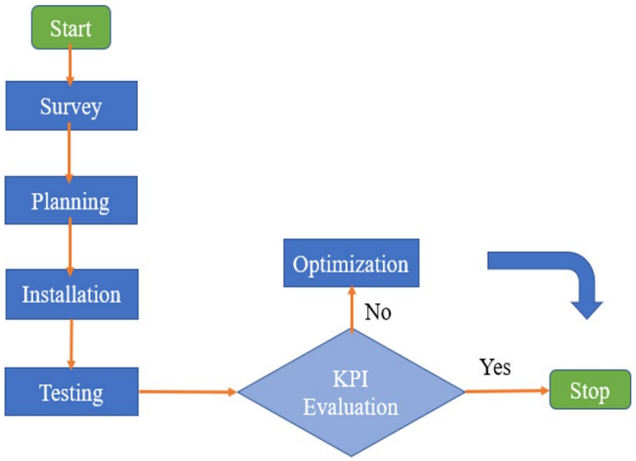


Fig. 4 Implementation flowchart

B. Walk Test Result

1) ***PCI-DL:*** There is 504 unique Physical cell index organized into 168 groups of three. PCI is thus uniquely defined by a number in the range of 0 to 167. We can know the coverage range of site PCI also. Based on legend PCI, the color of the cell is synchronized with the color coverage signal which shows us no cross feeder.

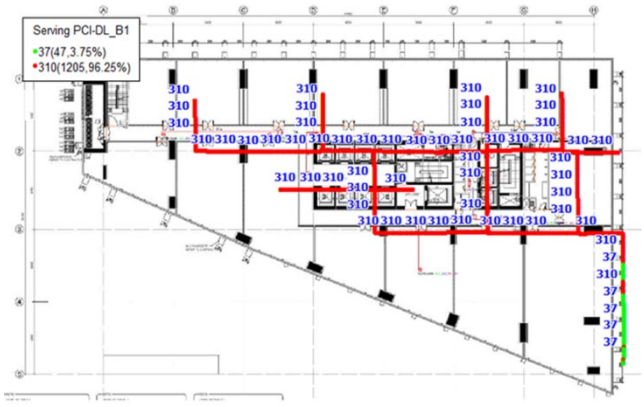


Fig. 5 PCI Result

2) ***RSRP-DL:*** After the site is on the site or UE to access, then we start doing KPI monitoring by collecting all log file for analysis. Based on reality some buildings are wall covered by a legend we collected data log file in the range from -80 to -60 is 81.55% excellent and -90 to -80 is 7.19% good.

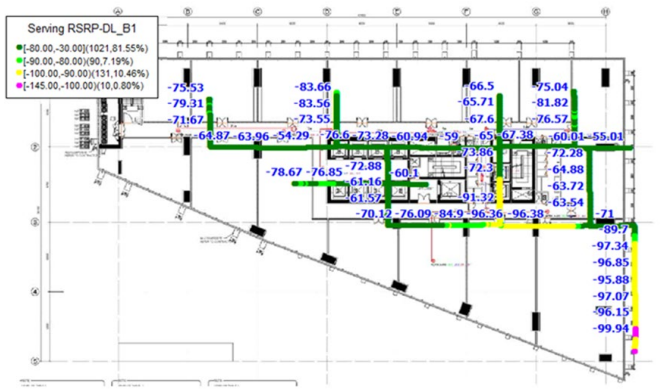


Fig. 6 RSRP Result

3) ***RSRQ-DL:*** After the site is on the site or UE to access, then we start doing KPI monitoring by collecting all log file for analysis. Based on reality some buildings are wall covered by a legend we collected data log file in the range from -10 to - 4 is 89.94% excellent and -15 to -10 is 9.11% good

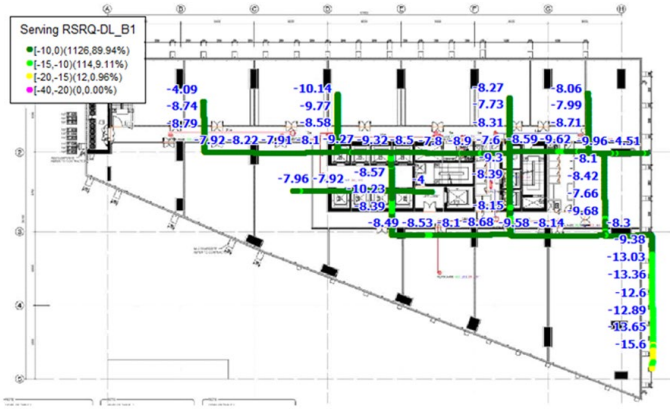


Fig. 7 RSRQ Result

4) ***Throughput Downlink:*** After the site is on the site or UE to access, then we start doing KPI monitoring by collecting all log file for analysis. Based on reality some buildings are wall covered by a legend we collected data log file in the range from 90M up is 0.5% excellent and 20M to 90M is 88.5% good.

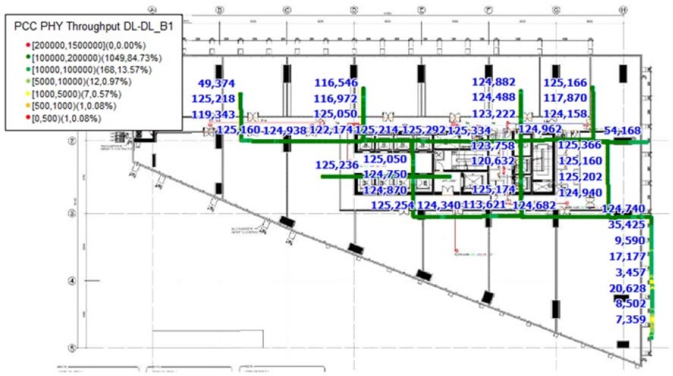


Fig. 8 Throughput Downlink Result

V. CONCLUSIONS

This project was finished successfully with 41 Floors that were built at some buildings and providing better 4G service in the whole building. After we installed, test, and optimize the whole building got good indoor coverage, and high capacity, strengthen the quality of mobile signals inside the building, and provide smooth and efficient wireless communication for users.

ACKNOWLEDGMENT

It is always a pleasure to remind me about those who always inspire me and give me the possibilities to complete my research paper program. Without their help from them, this whole project would not have been possible, and success on my paper internship. I would like to pay full respect and gratitude to my advisor. Mr. CHEA SOCHEAT who has been trying and working hard to explain, teach, and guide me in line to finish my report on my research paper internship.

REFERENCES

[01]. *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.

[02]. In-building solutions. php. Copyright © Unity Telecom Infrastructure Ltd., 2009.

[03]. A. Simonsson, M. Bergeron, J. Östergaard and C. Nizman, "*Real-life Indoor MIMO Performance with Ultra-compact LTE Nodes*", in Proc. of IEEE Vehicular Technology Conference Fall, 2015.

[04]. K. Hiltunen, "*Comparison of Different Network Densification Alternatives from the LTE Downlink Performance Point of View*" in Proc. of IEEE Vehicular Technology Conference Fall, 2011.

[05]. S. Zhou, M. Zhao, X. Xu, J. Wang, and Y. Yao, "*Distributed Wireless Communication System: A New Architecture for Future Public Wireless Access*", IEEE Commun. Mag., vol. 41, no. 3, pp. 108–113, Mar. 2003.

[06]. *Indoor-building-distributed antenna system (DAS)*. Wordpress.com, 2016.

[07]. IBS-training. Techpus Telcosys, 2013.

[08]. *Long-Term Evolution in Bullets*. Johnson, C. (2012).

[09]. *In-building wireless network design and deployment solution*. iBwave, 2013.

PageVis: Enhancing Facebook Data with a Visualization Tool

Heng Somnang[#1], Kor Sokchea[#2]

[#]*Department of Information Technology Engineering, FE, Royal University of Phnom Penh*

Phnom Penh, Cambodia

[1]heng.somnang.2018@rupp.edu.kh

[2]kor.sokchea@rupp.edu.kh

***Abstract*— Facebook has become a major social media platform that is growing exponentially because it is best for the connection between particular people, institutions, businesses, and organizations as well as government and non-government. With its popularity, many business owners ranging from small to big businesses start to shift their promotion campaign from traditional advertisement to Facebook advertising platform. In order to create a successful advertising campaign or provide personalized services, business owners need to understand their target customer by analysing user's data. Looking at raw data directly, it is not convenient. It needs tools and techniques that can turn data into graphic form. In this paper, we present a data visualization tool called PageVis that convert data fetching from Facebook through Facebook Graph API into wide variety of graphical representations. With this representation, it can help advertisers or business owners to understand their customer.**

***Keywords*— Facebook, Data Visualization, Facebook Graph API, PageVis, Visualization Tool**

I. INTRODUCTION

Facebook is a social media platform that is very popular in the world and also in Cambodia. It has around 2.934 billion monthly active users in July 2022 making it the most famous social media in the world [10]. With a large number of active users, various business owners has started to use social media platforms like Facebook for the promotions campaign instead of using traditional advertisement. The potential way to produce a successful advertising campaign and personalized service requirement, advertisers need to understand the user preference by analysing the data of users through provided data from Facebook Graph APIs. Facebook provides the graph API that allows analysers to fetch those data for analysing their target user. However, the data returning from Facebook Graph API is the raw data which is inconvenient for human to understand. Numeric representation is convenient for computer to read but not for human. Therefore, numeric representation needs to transform into graph representations by using visualization techniques in order to allow advertisers to gain insight into their target user.

In this paper, we will present a data visualization tool called **PageVis** that enable users to visualize various data from the Facebook page which aggregates structured and unstructured data sources including the data of the Facebook page insight, user data, and demographic data and converts them to virtual viewing through a dashboard of the software. With this tool, it gives a convenient way for a business owner to understand their audience and prepare strategies for the next advertising campaign in return for success.

II. LITERATURE REVIEW

The growth of data has been hailed as one of the greatest challenges of the 21st century due to the volume of data [1]. Data growth is defined as large, complex, high-velocity data sets [2]. However, big data brings new opportunities for research and new insights across a wide variety of fields [3] and by visually exploring and analysing the data. To capitalize on the data insight of social media platforms, visualization allows faster interpretation of the data, much of which is generated in real-time. Moreover, the goal of many businesses regarding data analysis is to recognize such patterns, emphasizing the need for data visualization to achieve strategic objectives. Visualizations can be likened to the front end of big data [4] and can be used to access and interpret the data, making the insights and trends more apparent. The Twitter analyser also uses the streamgraph to represent the behaviour of user interests' change over time [5]. Moreover, data from the tweet and retweets were plotted as a timeline with a line graph illustrating [6]. Twitter is one of the social media providing an API that could retrieve data geo-location coordinates and uses visualization techniques to plot data on the map by using OpenStreetMaps to compare the data for a period between different locations [7].

Common data visualization techniques to represent data are line charts, bar charts, area charts, graphs, and maps.

Line Charts are one of the most basic visualization techniques. They make the data more appealing and visualized. It shows the relationship between two patterns and compares several values at the same time interval. It is the most effective

approach when changes in a variable or variables need to be displayed [7].

Pie Chart is used to represent data in the form of a pie slice. The slice shows the amount of data. It is used to show the component's percentage of the whole. Moreover, Donut chat is one of the variations of pie charts [8].

Bar charts are referred to as column charts which make use of both horizontal and vertical bars. It could be used to compare items of a different group, but it is not very effective when the amount of data is very huge [8].

The map is used to plot the spatial data that is usually represented by Latitude, and Longitude. Individual locations are usually visualized as dots, and placed on a map according to a graphic coordination system [8].

Stream graph is used to show data values differing in median timeline. It represents the change of data over time. Thes method is helpful to visualize deeply in fast changing data sets. It would be the major feature of big data interpretation to make analyzers convenient to understand a mass of data to the graphic [9].

Additionally, there are many professional and commercial software which enable user to perform data visualization including Tableau, Microsoft Power BI, Qlik Sense, Looker and so on. Tableau is a tool that could visualize the interaction data visualization. It provides the potential range of visualization options. It is a very helpful tool since it is fast and flexible in using and also provides a wide variety of charts [9]. Microsoft Power BI would be the great cloud-based business analytics service. This tool is also flexible and persuasive [9]. And Plotly is built using python and Django framework. It is proficiency to perform analyzing and visualizing data. It allows users to create charts or dashboards for many purposes such as statistical charts, scientific charts, and multiple axes. Automatic grabs the data from the static images which known as "Web Plot Digitizer (WPD)" has been used by Plotly [9].

III. METHODOLOGY

1. FACEBOOK GRAPH API

.

Facebook Graph API is a primary way that Facebook provides to app developer to read and write to the Facebook social graph [11]. Graph API is a low-level HTTP-based API that could be utilized to query data, and fetch information related to posts and media upload [3]. Every content in the page including posts are restricted to the privacy setting of individuals. The secure access to the API is via access token. Access token is an open-source token generated for every Facebook user at the time of request. The token allows the app to interact with Graph API only until it is valid. Almost all Graph API endpoints require an access token of some kind, so each time you access an endpoint, your request may require one. Specific permissions have to be set to generate the access token as shown in Figure 1. These permissions are defined what kinds of data app can retrieve from the Graph API. The app can fetch only the data that is granted by the Facebook page owner. As shown in Fig. 1, we enable only **page_manage_posts**, **page_read_engagement**, **page_read_user_content**, and **page_show_list**. You can enable more permission by ticking on the permission you want to enable when you generate access tokens.

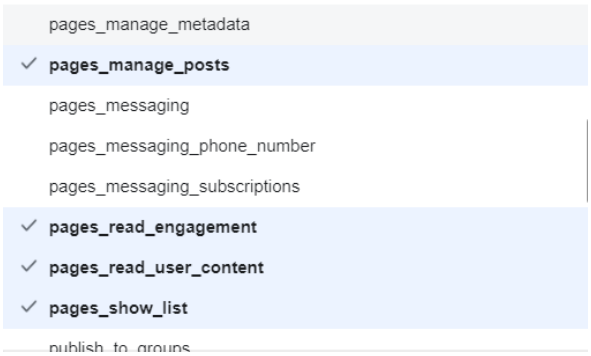


Fig. 1: Permission page

Facebook page could fetch data insight from Facebook Graph API unless that page consists of 100 or more likes. Data insight are not available for Facebook page that has page like lower than 100. The data that is fetched will update once every 24 hours. Moreover, demographic metrics, such as age, gender, and location, are returned if the data is at least 100 or more people. In Table I, you can see the requirement value for retrieving data from Graph API. You are not required to have an app to do experiment with Graph API. Facebook provides a tool called Facebook Graph API Explorer (FGAE) to allow developer to play around with the API. Developer can use FGAE to test, create, authenticate API calls, and debug responses. With FGAE, you can experiment what kind of data it will return from Graph API when you change the permissions.

TABLE I

REQUIREMENT VALUE FOR GETTING DATA

Type	Description
Access Tokens	A Page access token is requested by a person who can perform the ANALYZE task on the Page.
Features	Not applicable.
Permissions	read_insights, pages_read_engagement
Page Tasks	ANALYZE

2. FACEBOOK PAGE DATA FOR VISUALIZATION

There are a wide variety types of data on the Facebook page. However, the implementation in this paper was selected the specific types of data including **page view**, **page fan total**, **fan**

location, **fan age**, **page impression**, and **page fan removal**. These data follow the metric in the Table II which shows metric with detail explanation.

TABLE II
SELECTED DATA FROM FACEBOOK GRAPH API

Metric Name	Description	Values for 'period'
page_views_total	The number of times a Page has been viewed.	day, week, days_28
page_fans	The total number of people who have liked your Page.	day
page_fans_city	Aggregated Facebook location data, sorted by city	day
page_fans_gender_age	The number of likes of your Facebook Page.	day
page_impressions*	The number of times any content from your Page or about your Page entered a person's screen.	day, week, days_28
page_fan_removes_unique	Unlikes of your Page.	day, week, days_28

3. Request Data from Facebook Graph API

There are many ways to request the data from Facebook Graph API including HTTP, PHP SDK, Android SDK, iOS, and SDK. In this project, we chose the HTTP request as you see in Fig. 1.

```
GET v14.0/ {object-id}/insights/{metric} HTTP/1.1
```

Fig. 2: HTTP request

Example of requesting data from page_fans: after doing HTTP request to the Facebook graph API, it will response the result in the JSON format as shown in Fig. 3.

```
GET v14.0/me/insights/page_fans
```


```
{
  "data": [
    {
      "name": "page_fans",
      "period": "day",
      "values": [
        {
          "value": 9242,
          "end_time": "2022-06-27T07:00:00+0000"
        },
        {
          "value": 9443,
          "end_time": "2022-06-28T07:00:00+0000"
        }
      ],
      "title": "Lifetime Total Likes",
      "description": "Lifetime: The total number of people who have liked your P
      "id": "100724695778303/insights/page_fans/day"
    }
  ],
  "paging": {
    "previous": "https://graph.facebook.com/v14.0/100724695778303/insights?acces
```


Fig. 3: Data return from Facebook graph APIs

4. PREPARATION DATA FOR VISUALIZATION

The application was divided into two categories of the data. First is the overview of the Facebook page result that existed metrics such as page impression, page view, page fan, and page fan remove that will represent those data for a period of one month. Second is the demographic data that exist metric of data such as fan gender, age, and location.

The data that gets from Facebook Graph is returned as the JSON format. To visualize the overview of the Facebook page result we need to prepare the data in JSON format to the two lists that follow in Table III as the two-dimension that include a variable in one list and the value of the variable in one list of data, that will be easy to plot as the 2D graph.

TABLE III

DATA FORMAT

new_likes:	
["Jun-08",	[43,
"Jun-09",	50,
"Jun-10",	63,
"Jun-11",	47,
"Jun-12",	53,
"Jun-13",	72,
"Jun-14",	81,
"Jun-15",	64,
"Jun-16",	62,
"Jun-17",	74,
"Jun-18",	99,
"Jun-19",	100,
"Jun-20",	67,
"Jun-21",	60,
"Jun-22",	54,
"Jun-23",	158,
"Jun-24",	95,
"Jun-25",	95,
"Jun-26",	163,
"Jun-27",	165,
"Jun-28",	210,
"Jun-29",	298,
"Jun-30",	404,
"Jul-01",	533,
"Jul-02",	325,
"Jul-03",	304,
"Jul-04",	319,
"Jul-05",	339,
"Jul-06",	303,
"Jul-07"	353,
]	]

However, the structure of the data demographic is very similar to the data overview Facebook result that also converts those data in the format like shown in Table 3. It is an easy way to plot demographic data as the different 2D charts.

Moreover, the location data that returns from Facebook Graph API is the Name of the location (String). Therefore, the way that we handle this data is needed to convert a string of the location to latitude and longitude for plotting the data on the map. In this approach, we used the mapquest library to do the decoding. The access mapquest key API to do http request and append the values string of location Example: ("Phnom Penh, Cambodia")

Http request to mapquest as follows:

After converting, the data has generated the latitude and longitude of the location and needs to prepare that data to the format that shows in the Table IV.

TABLE IV

LOCATION DATA FORMAT

Location Name:	lat_long": [	
[	[	[
"Pailin, Cambodia",	12.87277,	102.63368,
"Prey Khm\u00ear, Cambodia",	11.48689,	104.85885,
"Phnom Srok, Cambodia",	11.56874,	104.92387,
"Suong, Cambodia",	11.88025,	105.68541,
"Kokong, Cambodia",	11.615815,	102.981519,
"Phumi Snuol, Cambodia",	11.52644,	104.9876,
"Phumi Banteay Neang, Cambodia",	13.46665,	103.01779,
"Srok Anlongweng, Cambodia",	11.56874,	104.92387,
"K\u00e2mp\u00f3ng Trach, Cambodia",	11.41674,	105.77377,
"Vientiane, Laos",	17.964099,	102.613371,
"Chhouk, Cambodia",	10.837861,	104.454052,

5. PAGEVIS

PageVis is the visualization tool created to visualize the data from the Facebook page. PageVis provide a dashboard for the user convenient to do visualization on the data that has been queried directly from the Facebook Graph API. There are two significant functions attached to the dashboard of the tools which are the input token and visualization area. To visualize the preferred page is required to input the valid token of the page. After the user input the valid token into the system, it automatically queries the data of the page from Facebook Graph API and stores the data in the raw data in the system. The next process is to convert fetching data into various graphical representations in the visualization area, we use the Matplot library. This library is used to turn the raw data represented into any chart and graph. Moreover, Folium library was used to visualize data on a map based on Latitude and Longitude. Matplotlib is a plotting library that provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits. Folium is a library to visualize data that's been manipulated in Python on an interactive leaflet map. It enables both the binding of data to map for choropleth visualization as well as passing rich vector/raster/html visualizations as a marker on the map. Additionally, the application clustered data into two categories for visualization. First, is the Overview Facebook page result, this part uses matplot that is manipulated with python, which plots the data to the different chart line charts and area charts. Visualizing a line or area chart is easy to compare the variation of data in the last thirty days. Second, for the demographic data, matplot was used to plot the bar chart, pie chart, and donut. In addition, geolocation data is used in the folium library to plot data of fans in the map based on the coordinate of latitude and longitude that was extracted in the previous process.

```
parameters = {
   "key": "api key"
   "location": "Preksandek, Cambodia"
}
respon=requests.get("http://www.mapquestapi.com/geocoding/v1/address", params=parameters)
```

Finally, PageVis was completely created and can visualize the Facebook page data in various types of data. The functionality of PageVis was separated into two categories and can visualize the flexibility graph depending on the attribute of the data as we have described in the previous point. The early part of PageVis was used to generate the overall Facebook data which includes the data of Total people view page, Pag Impression, and Number of people unfollowing or removing. Two options could visualize the total fan view page data as the line graph and area chart that will be visualized data change every data in last thirty days as shown in Fig. 4 and Fig. 5. Total page impression was represented as a line graph and area chart which is visualized data number of our page which reach to user screen every day in last thirty days as following in Fig. 7and Fig. 9. Number of people unfollowing is also represented as a line graph and an area chart since is convenient to understand the data change over time every day in last thirty days as following in Fig. 8 and Fig. 5. The second part of Pagevis could be to visualize the demographic data which include most of the personal information such as gender, age, and location of the fan who has liked or followed the page. On account of this, Pagevis has the proficiency to convert the data as raw text to various resilient graphs. The gender data of the user could represent in two options, the user can visualize this data with a pie chart, donut, bar chart that category data into three types of male, female, and undefine shown in Fig. 11, Fig. 13,and Fig. 14. Age data of user could visualize this data as a group of age,

and it has three options to choose for visualization as a bar chart and pie chart as following Fig. 15, Fig. 16, and Fig. 17. The location data could visualize as the bar graph and map as showing in Fig. 18, and Fig. 19. It can show the data of the page fan location where user live.

IV. CONCLUSION

The PageVis application has completed that have the ability to visualize various data from Facebook pages by extracting those data from Facebook Graph API. This software is really effective for the business owner or advertiser conveniently to understand the target of their customer through the data extracted as the picture with clear understanding since it helps encapsulate the data from the complicated data format converted to the graphic form. It also helps them easy to generate a daily report for their business.

However, there are some features that will improve the PageVis for more effectively tools that include more potential methods following:

- Extend more metric to visualization in this application
- Will be available to generate after visualization report
- Integrate with category content using a machine learning algorithm.

ACKNOWLEDGEMENT

I would like to express my gratitude thank to my Supervisor Kor Sokchea who is a professor at the Royal University of Phnom Penh for introducing this project to me and supervising the guide of this project and serving as an editor for this project.

REFERENCE

[1] A. Katal, M. Wazid and R. H. Goudar. (2013) "Big data: Issues, challenges, tools, and Good Practices," 2013 Sixth International Conference on Contemporary Computing (IC3). pp. 404-409,

[2] Seokyeon, K., et al.: “Big data visual analytics system for disease pattern analysis”, p 175 (2015)

[3] Resnyansky L.(2019). “Conceptual frameworks for social and cultural big data analytics: answering the epistemological challenge”. *Big Data Soc*. 6(1).

[4] Wang L, Wang G, Alexander CA. (2015). “Big data and visualization: methods, challenges and technology progress”. *Digit. Technol*.1 (1):33–38.

[5] Stojanovski, D., Dimitrovski,I., Madjrov,G. (2014). TWEETVIZ: TWITTER DATA VISAILIZATION.

[6] AI-Saqaf, W. (2016). “Mecodify: A tool for big data analysis & visualization with Twitter as a case study”.

[7] Sechelea, A., Huu, D, T., Zimos, E., & Deligiannis, N. (2016). “Twitter Data Clustering and Visualization.”

[8] Ganhi, P., Pruthi, J. (2020). “Data Visualization Techniques:Traditional Data to Big”.

[9] Ali SM., Gupta N., Nyak., GK., & Lenka RK (2016). “Big Data Visualization: Tools and challenges”

[10] https://datareportal.com/essential-facebook-stats#:~:text=Here's%20what%20the%20latest%20data,)%3A%202.934%20billion%20(July%202022)&text=Number%20of%20people%20who%20use,)%3A%201.968%20billion%20(July%202022)

[11] https://developers.facebook.com/docs/graph-api/

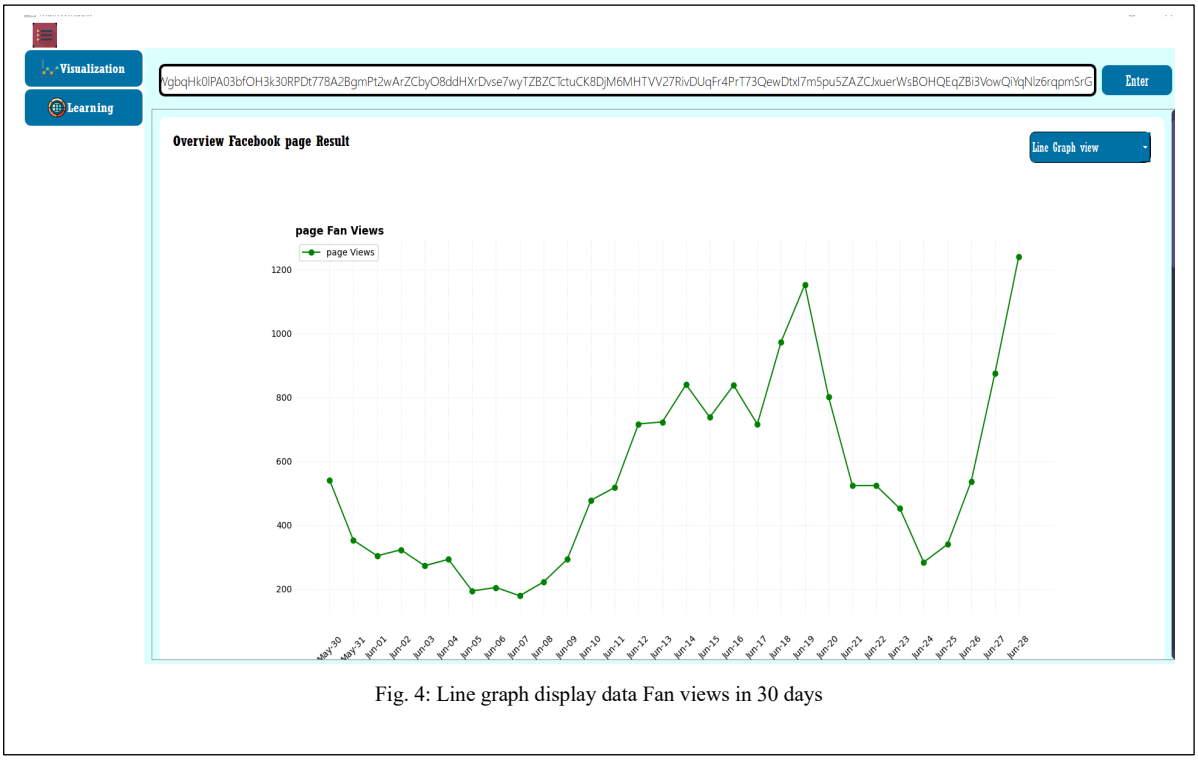


Fig. 4: Line graph display data Fan views in 30 days

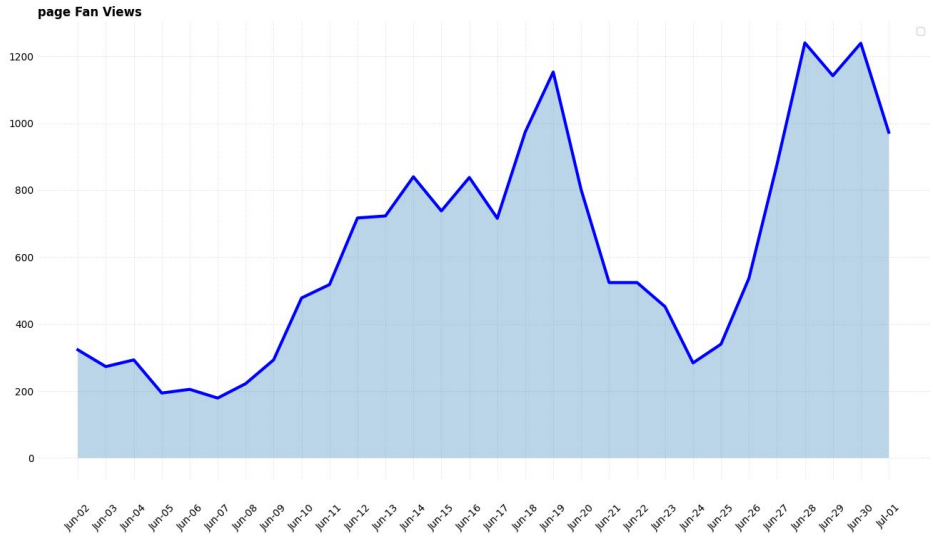


Fig. 6: Area graph display data Fan views in 30 days

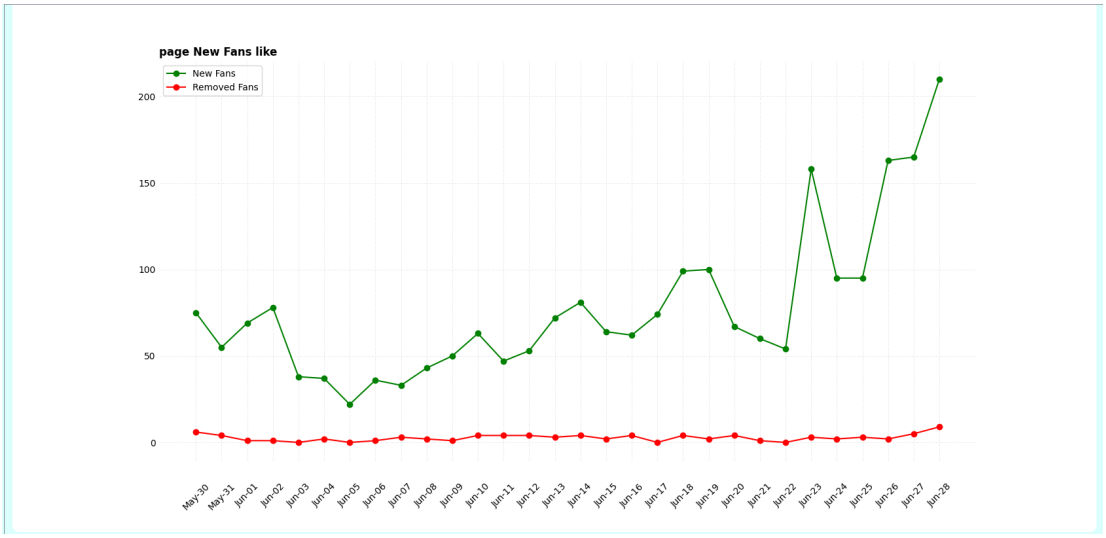


Fig. 5: Line graph display data New Fans like page and unlike in 30 days

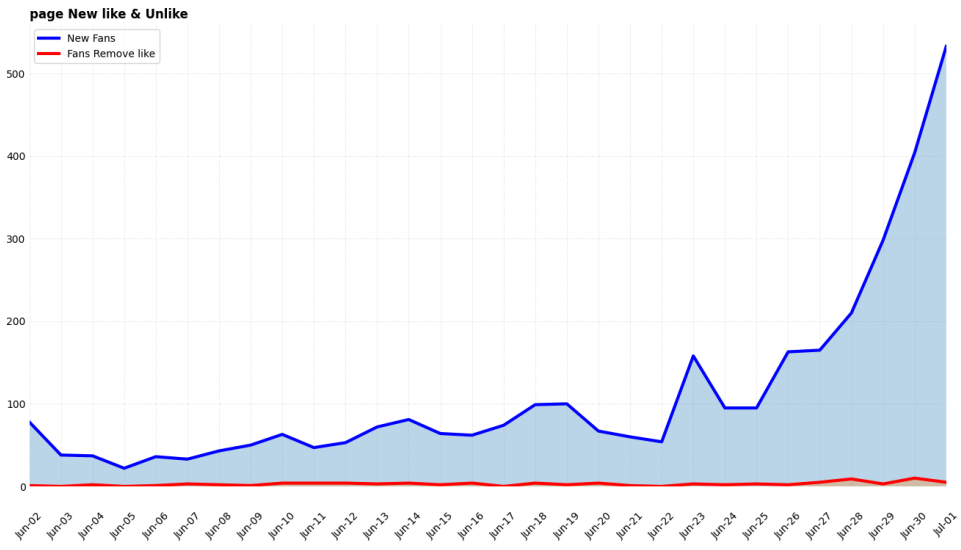


Fig. 8: area chart display data new like and unlike in 30 day

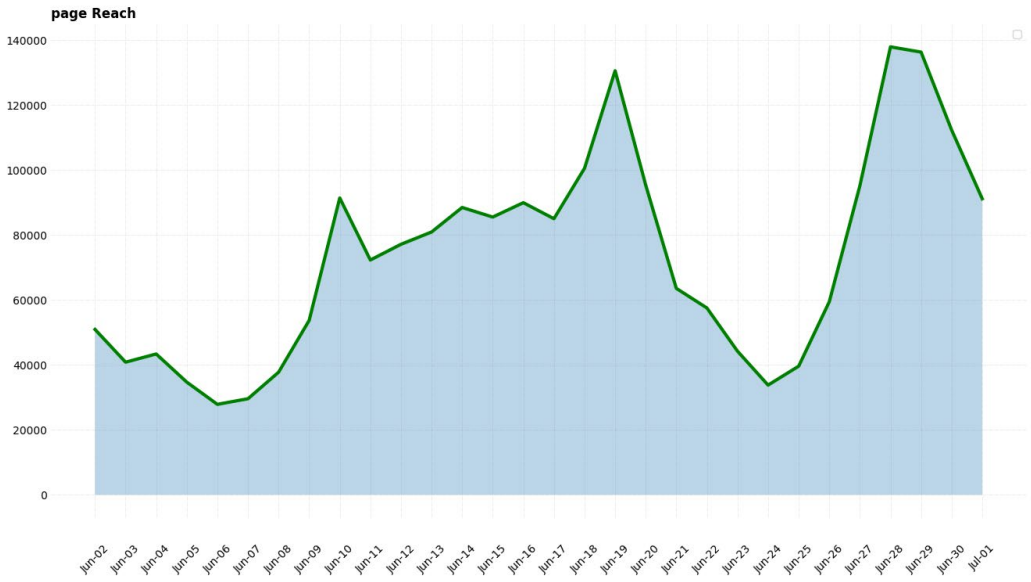


Fig. 7: Area chart display Page Impression in 30 days

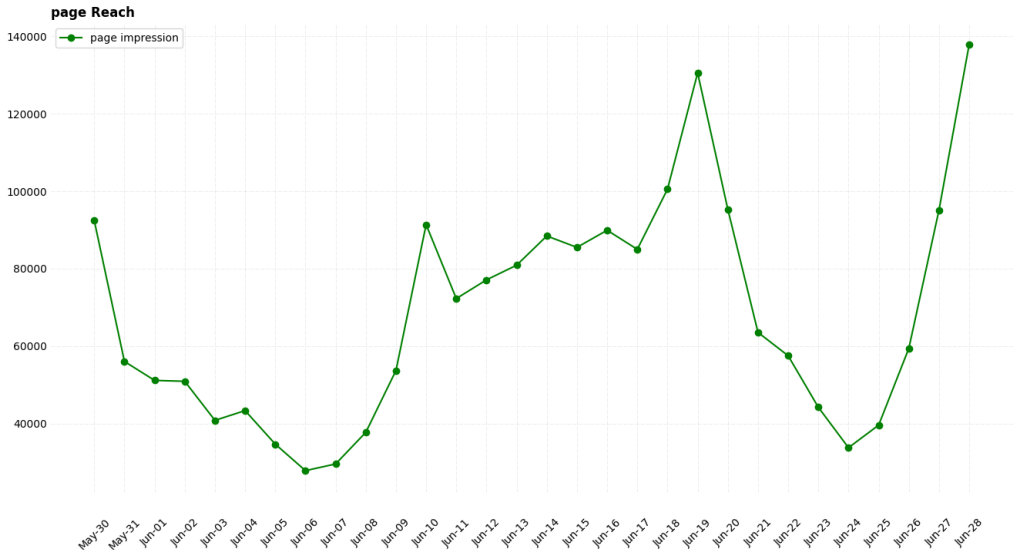


Fig. 9: Line graph display Page Impression in 30 days

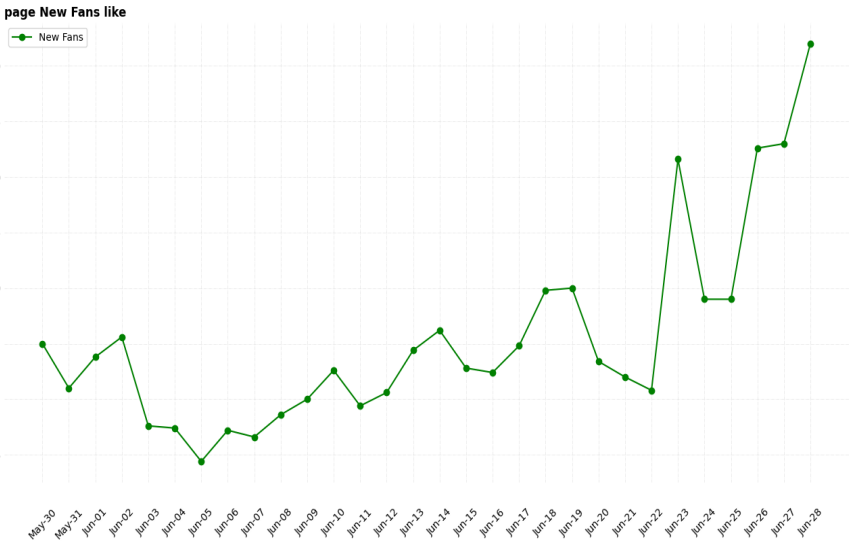


Fig. 10: Line Graph display new Fan like page in 30 days

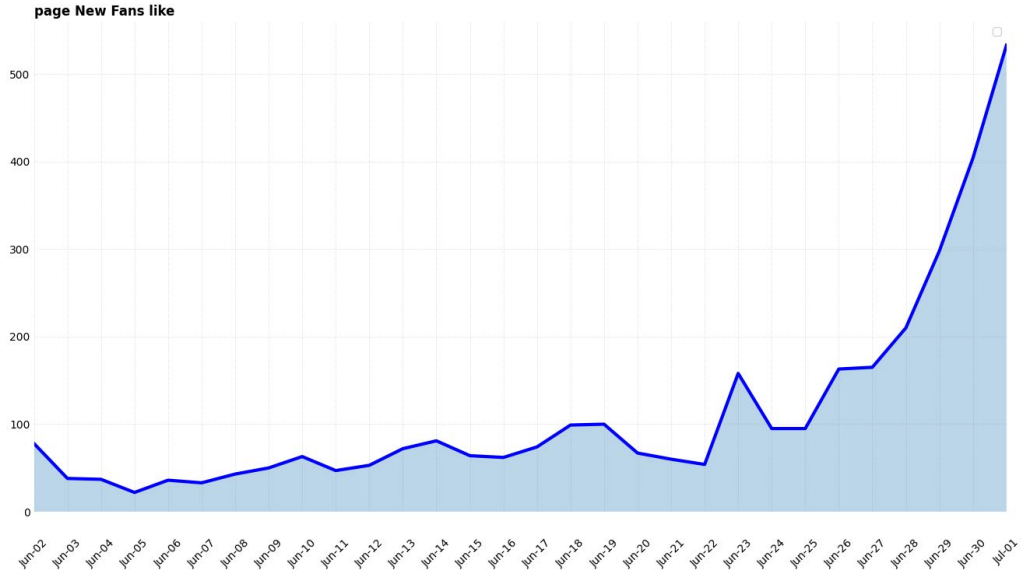


Fig. 12: Area chart display new Fan like page in 30 days

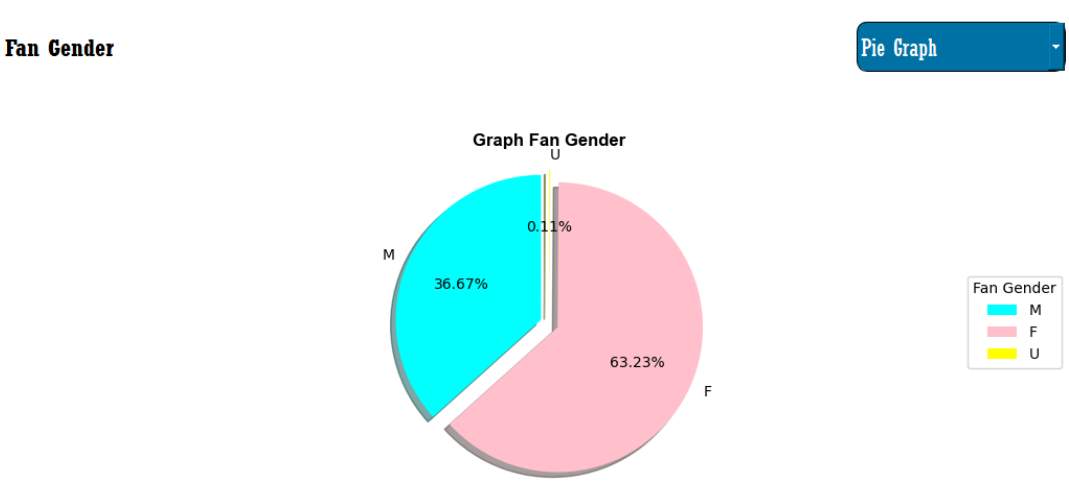


Fig. 11: Pie chart display gender data of page fan

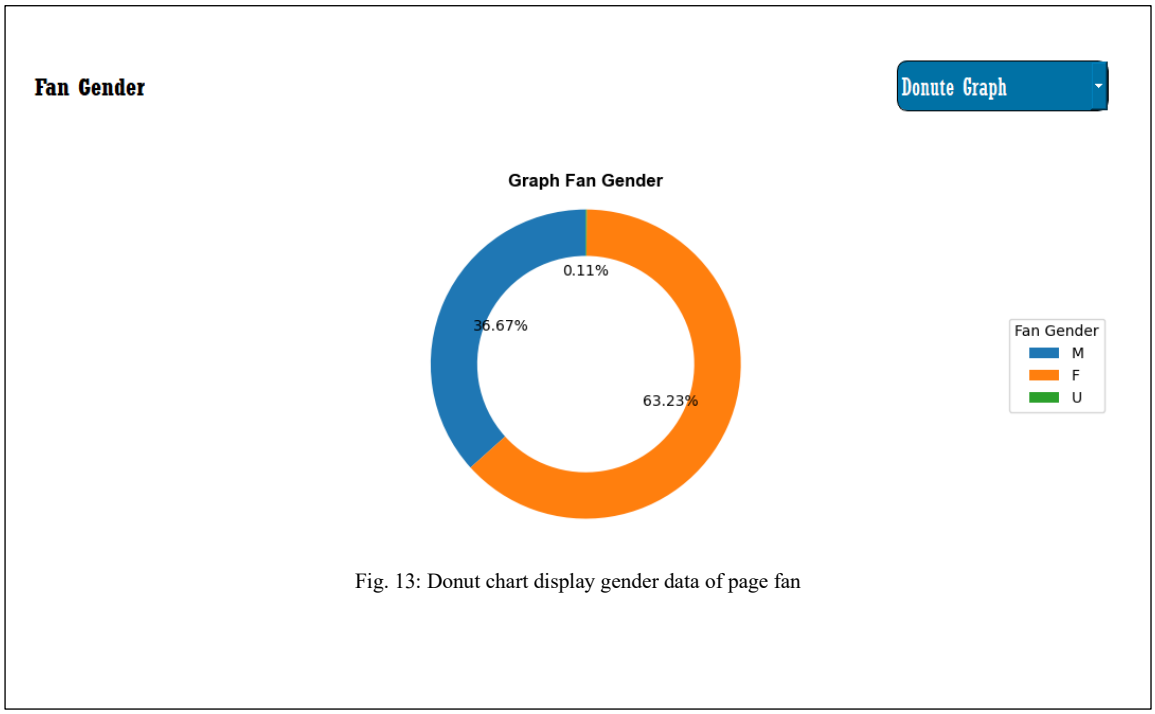


Fig. 13: Donut chart display gender data of page fan

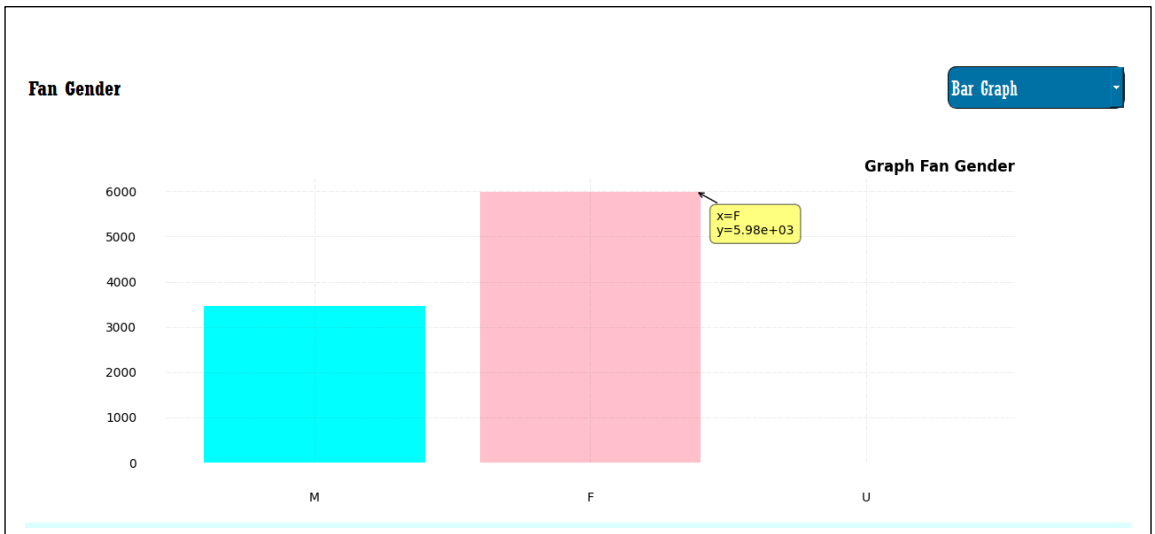


Fig. 14: Bar chart display gender data of page fan

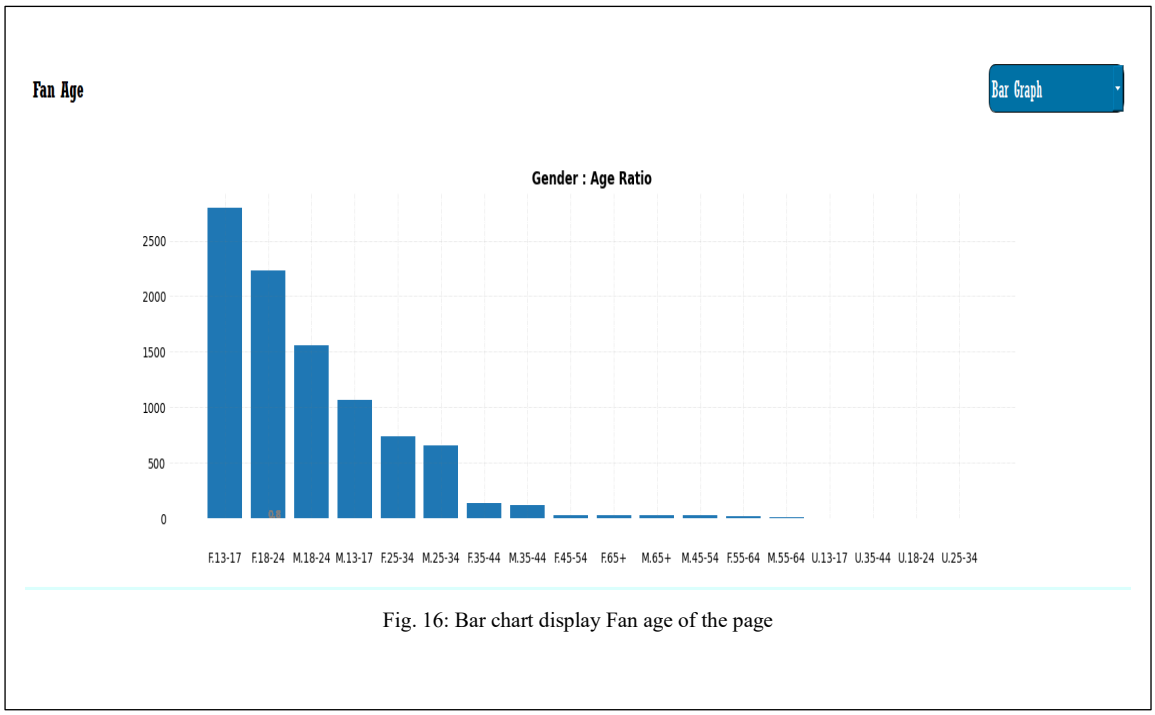


Fig. 16: Bar chart display Fan age of the page

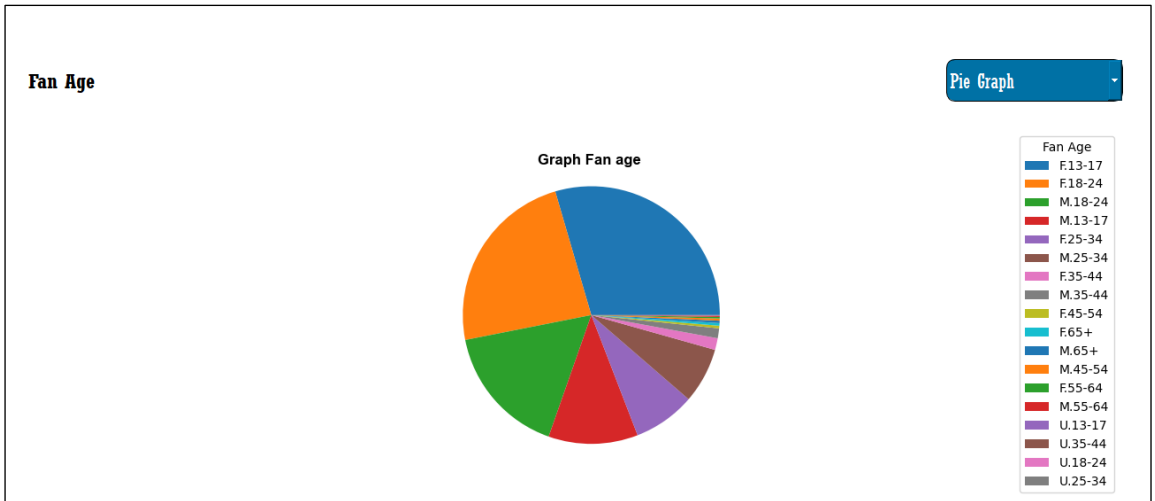


Fig. 15: Pie chart display Fan age of the page

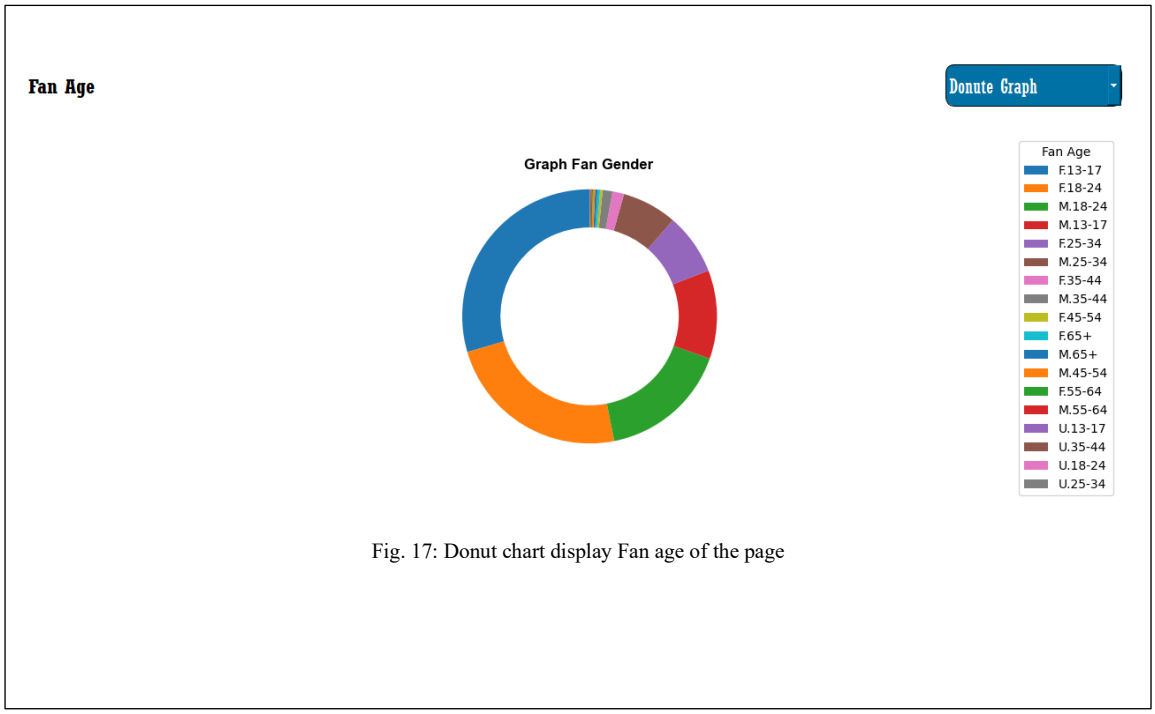


Fig. 17: Donut chart display Fan age of the page

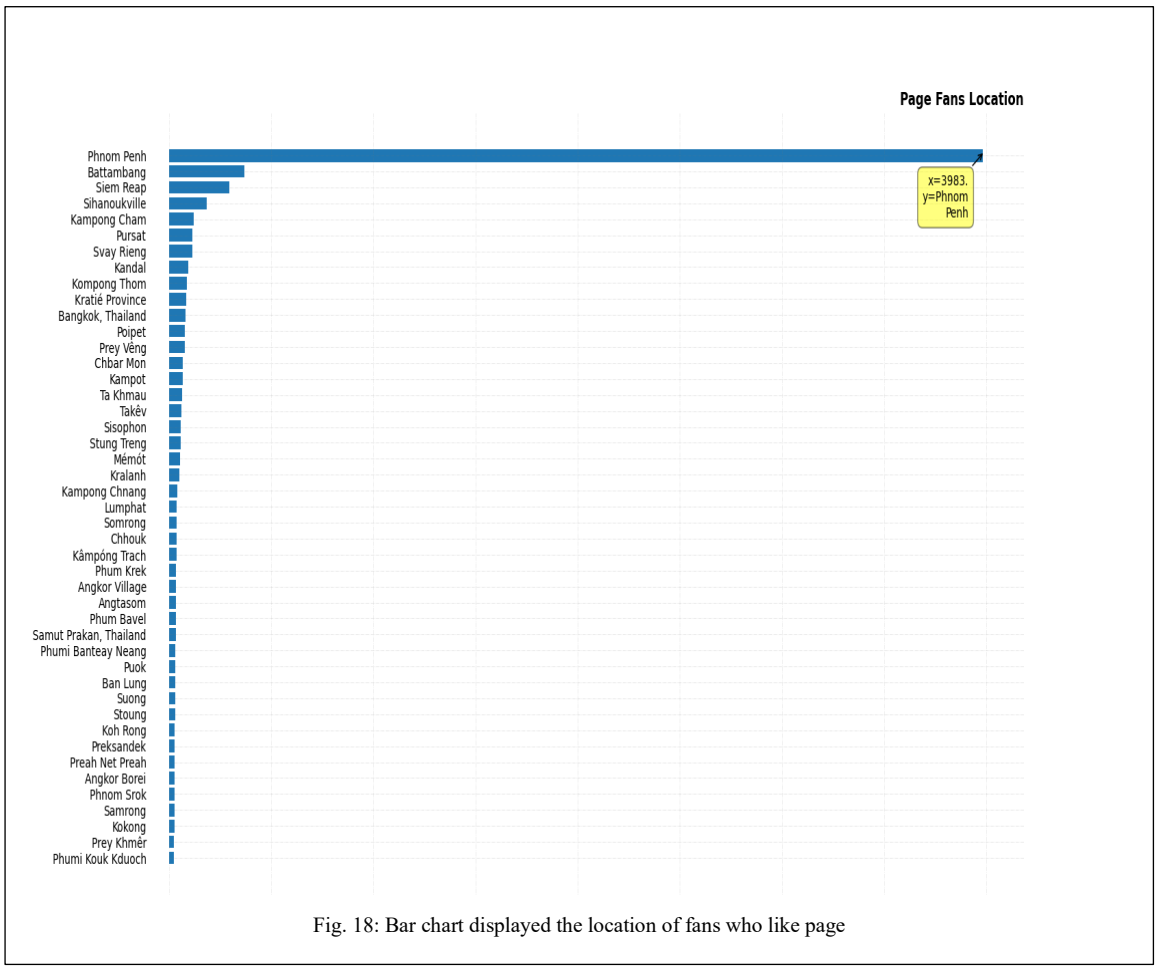


Fig. 18: Bar chart displayed the location of fans who like page

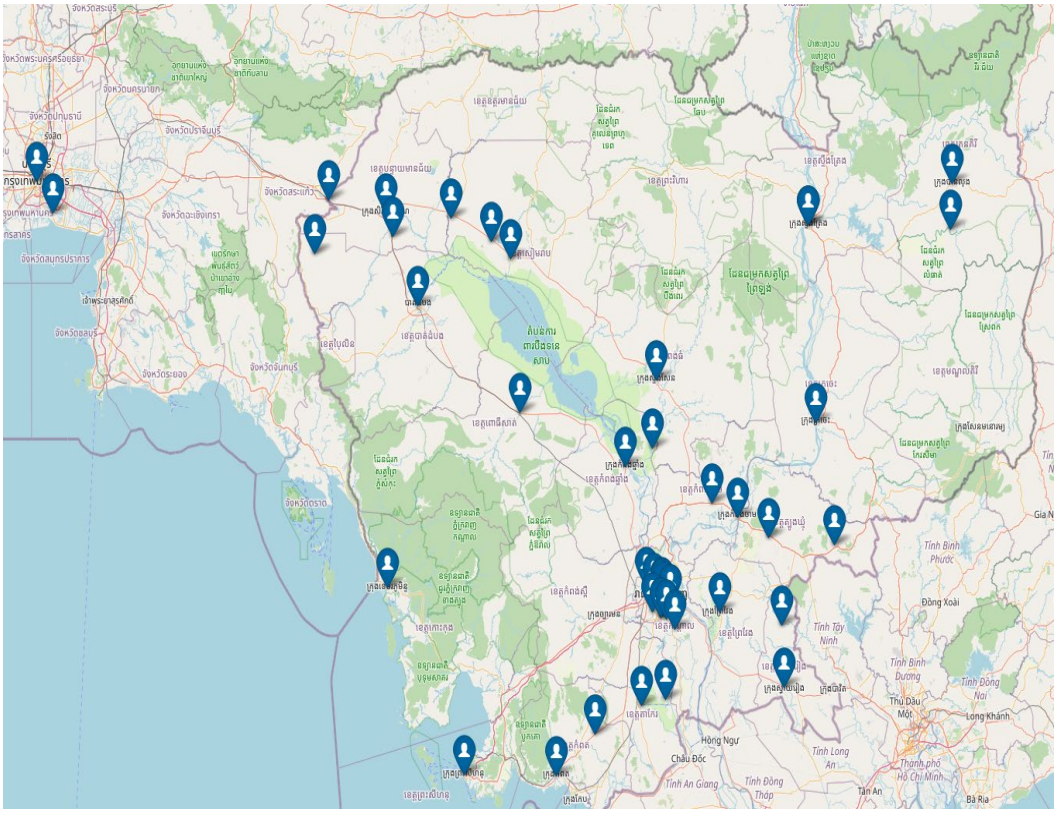


Fig. 19: Map displayed the location of fans who like page

Predicting Facebook Posts category using Multinomial Naïve Bayes Classifier

Manitou PHON[#1], Sokchea KOR[#2]

[#]*Department of Information Technology Engineering, FE, Royal University of Phnom Penh*
Phnom Penh, Cambodia

[1]phon.manitou.2018@rupp.edu.kh

[2]kor.sokchea@rupp.edu.kh

***Abstract*— Text classification is a process of categorising texts into organised groups. It's the fundamental part of Natural Language Processing. By doing text classification, it helps us with making use of the unstructured texts and transform them into a more understandable information which grants us some valuable insights. Unstructured data will have some noises or unnecessary words especially the data that has been extracted through Facebook GraphAPI. In this study, our objective is to apply word representation language model called bag of words in combination with Multinomial Naïve Bayes Classifier to predict Facebook Posts category. Due to noise in data, some text pre-processing techniques has been used to remove the unwanted noise or words before performing classification. In this experiment, we have achieved an overall F-1 score of 0.75.**

***Keywords*— text classification, Multinomial naïve Bayes, unsupervised learning, machine learning, Facebook.**

I. Introduction

This document describes about the process of applying a text classifying technique to classify Facebook posts categories. Text classification is a technique to categorise a text document dataset into small and different classes provided that each class inherits different properties, functionalities and features from each other. An illustration diagram about Text Classification can be shown as below.

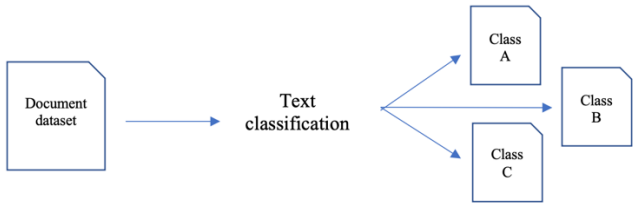


Fig. 1 Text classification diagram

A. Facebook

Facebook is one of the strong standing social media platform on the internet. It was the foundation and the base of the other social media platform. Sharing information is the core part of what we define a social media platform. While Facebook is being one of the most used, it's very useful to understand each posts information as much as possible to understand the demographics and target of those posts.

B. Aim and Objectives

The aim of this study is to provide news classification for Facebook Pages along with engagement score for insights.

- The first objective is to classify the content from each Facebook posts
- The second and last objective is to calculate the user engagement scores with each post.

II. Literature Review

"Over the last few years, classification news headlines have been an area of research including, classification of emotions, classification of financial news, headlines classification with N-gram model, automated categorization of news headlines classification of news headlines with SVM, mining of emotions from headlines of news, and Twitter classification of news from short news headlines. It is very important to have a proper news headlines categorization in our lives. Text classification is a method of allocating predefined categories to a text in conformity with its contents. A well-categorised dataset of news has to be utilised for exploration such as prediction of the stock market, news categorization and trading system, and news-oriented stock trend prediction."[8]

There are numerous text classifiers like Support Vector Machines, KNN, decision trees, Neural Network Classifier, gradient boosting etc. According to a comparative study of Automatic Semantic Categorisation of News Headlines using Ensemble Machine Learning [5] , It is witnessed that Multinomial Naïve Bayes has the highest accuracy for news text classification. The result comparison from that study is shown as follow.

TABLE I
RESULTS COMPARISON

Classifier	Class	Precision	Recall	F1-Score
KNN	Travel	0.82	0.82	0.82
	Style & Beauty	0.86	0.87	0.86
	Parenting	0.86	0.85	0.85
SVM	Travel	0.87	0.90	0.88
	Style & Beauty	0.92	0.89	0.90
	Parenting	0.91	0.90	0.90
Multinomial Naïve Bayes	Travel	0.87	0.91	0.81
	Style & Beauty	0.89	89	0.89
	Parenting	0.91	0.90	0.90
Gradient Boosting	Travel	0.88	0.87	0.87
	Style & Beauty	0.92	0.87	0.89
	Parenting	0.86	0.91	0.88

Moreover, since Naïve Bayes Classifiers assume that the value of a particular feature is independent of the value of any other feature, it's very efficient.

III. METHODOLOGY

A. Data Collection

1) Facebook Page data collection

We use Facebook graph's API to extract each Facebook page's post data. The following will be discussing about how we are going to extracting data. According to Facebook graph API document, we could utilise the insight endpoint. The API would then return a JSON object which consists of *Keys* and *Values*. At the time this document is published, we're using Facebook's GraphAPI v14.0. We can then programmatically loop and extract the required data such as the "*message"*, which is the title (caption) of the post, "*posts_clicks", "post_negative_feedback",* "*shares", "reactions"* and "*comments".*

2) Data scraping for testing

Since we don't own a page where we posted news related post in English, we decided to scrape public posts from a local news page (*Phnom Penh)*. This document will describe the use of python to scrape Facebook posts. We would scrape *Facebook Post Captions* by scraping "Text" as describe in the library's document. Since Facebook has limitation for posts scraping and would ban the scraper IP address for 24 hours, we programmed the scraper to scrape between a random interval between 30 and 90 seconds.

B. Data Processing.

The classification consists of 3 main procedures. The diagram of model can be found in the proposed model below.

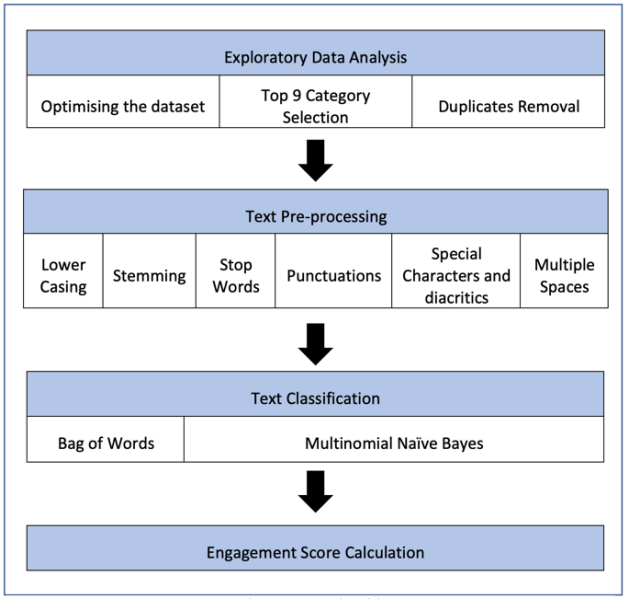


Fig. 2 Proposed Model

1. Exploratory data analysis

The idea of this procedure is to optimise the selected dataset for processing as clean as possible. In this report, we use categorised news dataset from Kaggle [9]. The dataset provides category, headline, authors, link, short description and publish date. There are so many

The original data set consists of the following categories:

TABLE III
THE ORIGINAL DATASET CATEGORY

Category Name	Numbers of category
POLITICS	32739
WELLNESS	17827
ENTERTAINMENT	16058
TRAVEL	9887
STYLE & BEAUTY	9649
PARENTING	8677
HEALTHY LIVING	6694
QUEER VOICES	6314
FOOD & DRINK	6226
BUSINESS	5937
COMEDY	5175
SPORTS	4884
BLACK VOICES	4528
HOME & LIVING	4195
PARENTS	3955
THE WORLDPOST	3664
WEDDINGS	3651
WOMEN	3490
IMPACT	3459
DIVORCE	3426
CRIME	3405
MEDIA	2815
WEIRD NEWS	2670
GREEN	2622
WORLDPOST	2579
RELIGION	2556
STYLE	2254
SCIENCE	2178
WORLD NEWS	2177
TASTE	2096
TECH	2082
MONEY	1707
ARTS	1509
FIFTY	1401
GOOD NEWS	1398
ARTS & CULTURE	1339
ENVIRONMENT	1323
COLLEGE	1144
LATINO VOICES	1129
CULTURE & ARTS	1030
EDUCATION	1004

As we can see, we could optimise the news dataset by selecting only the category, headline concatenated with its short description.

There are a total of 41 categories of news headlines within the selected dataset. However, we can improve the accuracy of the prediction by optimising and selecting the top 10 categories for processing.

we can generate top 9 category for our modified dataset as the following:

TABLE IIIV
MODIFIED DATASET CATEGORY

Category Name	Numbers of category
POLITICS	32739
ENTERTAINMENT	25111
LIFE STYLE	18597
HEALTH & WELLNESS	17827
WOMAN & CHILD	12167
BUSINESS	7644
SPORTS	4884
SCIENCE & TECH	4260
EDUCATION	2148

2. Text pre-processing

The main goal of pre-processing text is to clean every text. To achieve this, we would like to apply the following methods:

- replace Not assigned values with empty spaces.
- remove blocks of digits.
- remove all punctuations such as
- !”#$%&’()*+,-./:;<=>?@[]^_`{|}~).
- remove all diacritics and accents from any word and characters.
- remove all instances of all English stop-words. In this report, we will be using NLTK Error! Reference source not found. stop-words which contains 179 of stop-words.
- remove any extra white-spaces

3. Text classification

3.1. Bag of word model.

In order to mathematically represent each word as vectors, we would utilise the concept *bag of word* as the word representation model. This model will represent texts as the bag of its word. For example, based on the 2 following documents, we can construct each word in a list as:
(1) John likes to watch movies. Mary likes movies too.
(2) Mary also likes to watch football games.

would be converted to:

"John","likes","to","watch","movies","Mary","likes","movies","too"

"Mary","also","likes","to","watch","football","games"

The bag of word model representation would be represented in JSON object as:

BoW1 = {"John":1,"likes":2,"to":1,"watch":1,"movies":2,"Mary":1,"too":1};

BoW2 = {"Mary":1,"also":1,"likes":1,"to":1,"watch":1,"football":1,"games":1};

If we combine both BoW1 and BoW2, we would get BoW3 as the union set of BoW1 with BoW2:

BoW3 = {"John":1,"likes":3,"to":2,"watch":2,"movies":2,"Mary":2,"too":1,"also":1,"football":1,"games":1};

Thus, BoW2 is the union of BoW1 and BoW2

$$BoW3 = BoW1 \biguplus BoW2$$

Multinomial Naïve Bayes

Naive Bayes is an abstract concept of conditional probability, which can be represented by $x = (x_1, \dots, xn)$ where n is the number of features. It assigns to the instance probabilities of $p(C_k \mid x_1 \dots, x_n)$for each K possible outcomes or classes C_k

By applying Bayes's Theorem, the conditional probability can be calculated as

$$p(C_k \mid x) = \frac{p(C_k)p(x \mid C_k)}{p(x)}$$

With a Multinomial event model, samples (feature vectors) represent the frequencies with which certain events have been generated by a multinomial $(p_1, \dots, p_n)$, where p_i is the probability that event i occurs

A feature vector, $x = (x_1, \dots, x_n)$ is a histogram where x_i is the number of times even i was observed in a particular instance. The likelihood of observing a histogram x is given by

$$p(\mathbf{x} \mid C_k) = \frac{(\sum_{i=1}^{n} x_i)!}{\prod_{i=1}^{n} x_i!} \prod_{i=1}^{n} p_{ki}{}^{x_i}$$

Engagement score calculation

To rank each and every post, we need to calculate how engaged did the audience interacted with each of our posts. We measured the engagement score by assuming the following equation:

$$Ps = (P_{tr} * 1.1) + (P_{cm} * 1.4) + (P_{sh} * 1.2) + P_{ck}$$

Where:

- Ps: Post engagement Score
- P_{tr}: Post Total Reactions (Likes, Heart, Care, Wow, Ha-ha, Sad and Angry)
- P_{cm}: Post total Comments
- P_{sh}: Post total shares

The engagement score shows which posts had the most interactions and would be interested for the audience.

IV. Results

This chapter presents the algorithm's result and the application user interface. The first part of this result will be discussed about the algorithm's result.

A. User Interface

In this document, we use PyQT to demonstrate the ideas. The following figures are the result of the UI.

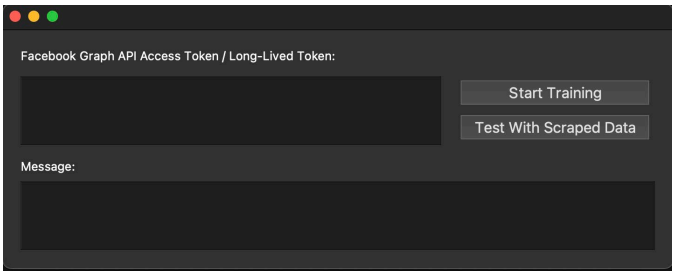


Fig. 3 Main UI window

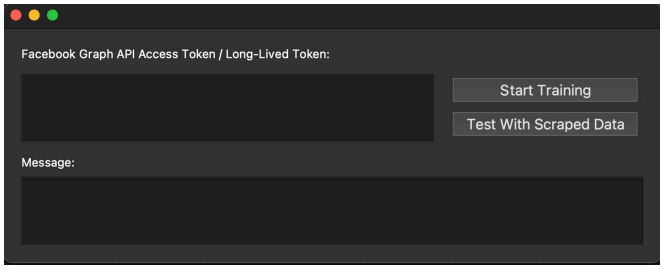


Fig. 4 UI after filling the token input box

After clicking the Start processing button, the program will start collecting Posts from Facebook and After it's finished, it will start the data processing procedures. The following will show the result after the procedures are finished.

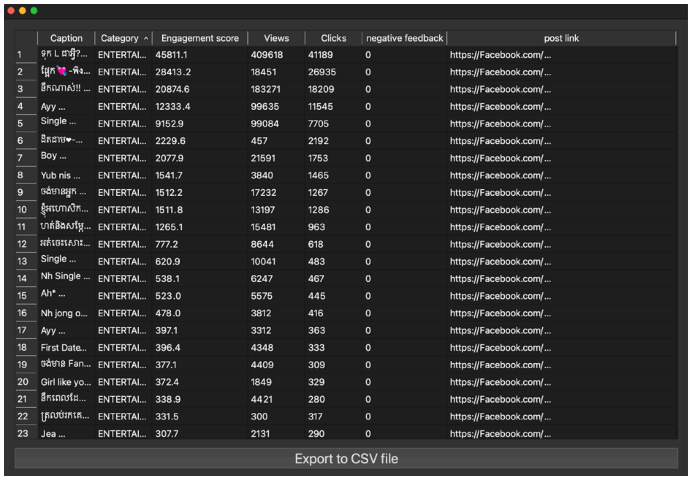


Fig. 5 The result after the procedures are finished

shows the expected columns. Since we don't own a page related to news, we will be going to use the scraped dataset from scraping to demonstrate the predicting capabilities. By clicking on the Test with Scraped data button, we will use the scraped dataset instead.

Fig 6 The result after using the scrapped data

To save the current table result, we can click the Export to CSV file button and we would get the save CSV file dialogue as with the following figure.

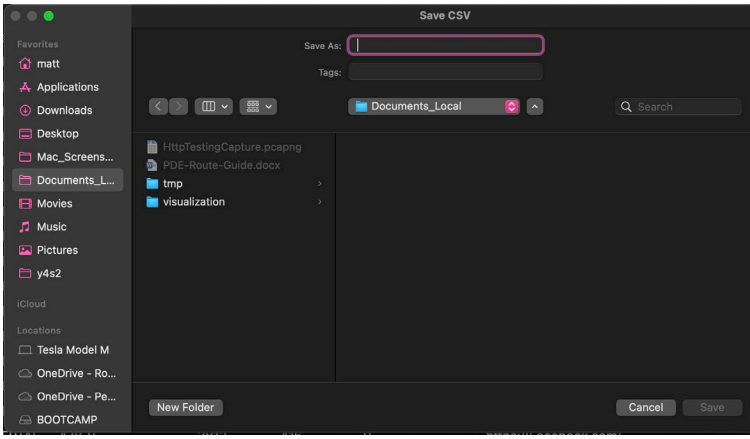


Fig. 7 Save the result as CSV

B. *Classifier results and discussion*

In this section, we will investigate a few ways to increase the accuracy of the classifier. The classification report is shown as in figure 8, as we can see, the class education has the lowest recall since there was not enough data to extract its feature to make the prediction.

TABLE V
CLASSIFIER REPORT

	Precision	Recall	F1-Score	Support
Business	0.66	0.55	0.6	2250
Education	0.73	0.14	0.24	610
Entertainment	0.78	0.78	0.78	7586
Health & Wellness	0.59	0.87	0.7	5303
Life Style	0.78	0.58	0.66	5609
Politics	0.82	0.91	0.86	9949
Science & Tech	0.81	0.35	0.49	1223
Sports	0.89	0.6	0.71	1434
Woman & Child	0.63	0.62	0.62	3650
accuracy			**0.74**	**37614**
Macro Avg	**0.74**	**0.6**	**0.63**	**35364**
Weighted Avg	**0.75**	**0.74**	**0.73**	**72368**

Politics	90.97%	3.57%	0.34%	1.40%	1.31%	1.90%	0.21%	0.13%	0.17%
Entertainment	8.03%	77.72%	5.50%	2.99%	4.19%	0.79%	0.38%	0.36%	0.04%
Life Style	3.48%	5.69%	57.91%	27.72%	3.67%	1.07%	0.23%	0.21%	0.02%
Health&Wellness	1.72%	2.41%	2.73%	86.65%	4.94%	1.07%	0.19%	0.26%	0.02%
Woman&Child	6.96%	9.40%	4.41%	15.53%	61.59%	1.29%	0.52%	0.14%	0.16%
Business	17.33%	4.44%	2.93%	13.91%	4.76%	54.98%	0.44%	1.07%	0.13%
Sports	11.79%	17.29%	2.51%	3.35%	4.32%	0.70%	59.83%	0.21%	0.00%
Science&Tech	10.79%	12.84%	4.74%	17.17%	6.87%	11.61%	0.57%	35.32%	0.08%
Education	30.00%	5.08%	0.82%	16.56%	20.98%	11.15%	0.33%	0.66%	14.43%
	Politics	Entertainment	Life Style	Health &Wellness	Woman &Child	Business	Sports	Science &Tech	Education

Fig. 8 Confusion Matrix

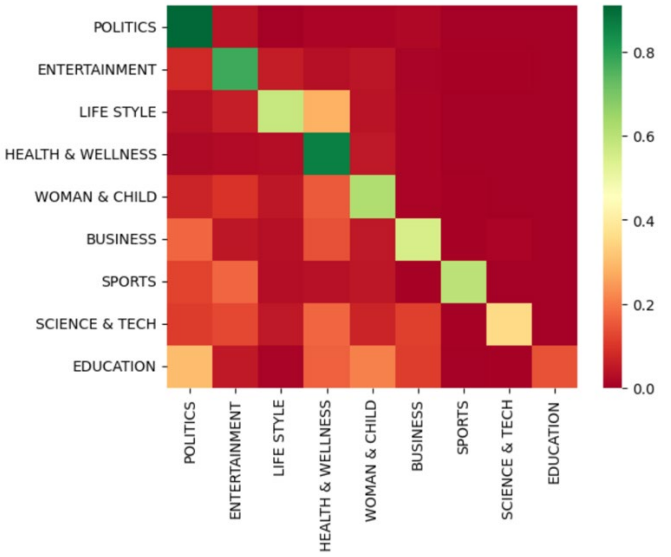


Fig 9 Confusion Matrix heat-map

If we take a look at figure 9, the heat-map shows that the classifier still has some confusion between **politics** and **education** since politics has the most data for processing while education have the least data. We could improve the dataset and implement some unsupervised learning before applying Multinomial Naïve Bayes classifier to eliminate most of the confusions between classes.

The overall accuracy of the classifier is 0.7352 which means this model would predict inputs with 73.52% accuracy.

The result shows that the dataset model from processing was not perfect since there are some major and minor class bias. However, with better a better dataset and a more distinct categories data model, we could achieve a more accurate prediction.

V. Conclusion and Future Works

This chapter presents the algorithm's result and the application user interface. The first part of this result will be discussed about the algorithm's result.

The object of the task was to predict which category have the most engagement for that Facebook page audiences. After the implementation, we got a great idea of which post had the highest engagements. Now, we have an idea of which type of post should the Facebook page uses to attract their audiences. Even if the classifier's overall accuracy is not perfect, we could still have a glimpse of idea on which category belongs to the post with highest engagements from the page audience. The future work for this model will involve implementing these features:

- Improve the accuracy by implementing unsupervised learning to pre-classify the classes.
- Gather more data for the prepossessing dataset
- Apply self-learning techniques
- Automatic prediction with browser extensions.

References

[1] Parveen, H., & Pandey, S. (2016, July). Sentiment analysis on Twitter Data-set using Naive Bayes algorithm. In 2016 2nd international conference on applied and theoretical computing and communication technology (iCATccT) (pp. 416-419). IEEE. https://ieeexplore.ieee.org/abstract/document/7912034

[2] Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: a statistical framework. International journal of machine learning and cybernetics, 1(1), 43-52. https://link.springer.com/article/10.1007/s13042-010-0001-0

[3] Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks, 5(1), 7-16. https://www.researchgate.net/profile/Vijayarani-Mohan/publication/339529230_Preprocessing_Techniques_for_Text_Mining_-_An_Overview/links/5e57a0f7299bf1bdb83e7505/Preprocessing-Techniques-for-Text-Mining-An-Overview.pdf

[4] Text News Classification System using Naïve Bayes Classifier http://ijoes.vidyapublications.com/paper/Vol13/39-Vol13.pdf

[5] Automatic Semantic Categorization of News Headlines using Ensemble Machine Learning: A Comparative Study https://pdfs.semanticscholar.org/145f/f417d3eae51be4e710f2489359a1d122eacf.pdf

[6] Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. Journal of advances in information technology, 1(1), 4-20. https://www.academia.edu/download/38147943/jait0101.pdf#page=6

[7] Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. (2004, December). Multinomial naive bayes for text categorization revisited. In Australasian Joint Conference on Artificial Intelligence (pp. 488-499). Springer, Berlin, Heidelberg. https://link.springer.com/chapter/10.1007/978-3-540-30549-1_43

[8] Bashir, E., & Luštrek, M. (2021). Self Learning of News Category Using AI Techniques. In Intelligent Environments 2021: Workshop Proceedings of the 17th International Conference on Intelligent Environments (Vol. 29, p. 167). IOS Press. https://ebooks.iospress.nl/pdf/doi/10.3233/AISE210094

[9] Misra, Rishabh (2018). News Category Dataset. Kaggle DOI: 10.13140/RG.2.2.20331.18729 https://www.kaggle.com/datasets/rmisra/news-category-dataset

[10] Facebook Graph API https://developers.facebook.com/docs/graph-api/

[11] GitHub: kevinzg/facebook-scraper https://github.com/kevinzg/facebook-scraper

A Comprehensive Survey on Recommendation System

Nuth Vireak[#1], Tan Bunchhay[#2], Kor Sokchea[*3]

[#]*Department of Information Technology Engineering, Faculty of Engineering, Royal University of Phnom Penh*
Phnom Penh, Cambodia

[1]2820.nuth.vireak@rupp.edu.kh

[2]2820.tan.bunchhay@rupp.edu.kh

[*]*Corresponding author: Kor Sokchea*

[3]kor.sokchea@rupp.edu.kh

Abstract— **Recommendation system has received much attention in the past decade due to the increase in demand for e-commerce and online services. It is a software tool and algorithms providing suggestions to users based on their interests to help them find new products and services they might be interested in by filtering out information that is not relevant to their specific tastes or preferences [1]. Big tech and e-commerce companies use this algorithm to deliver great personalized services and appealing user experiences that enable companies to retain their customers and increase sales. In October 2006, Netflix held a competition called "The Netflix Price" to develop an algorithm that could improve their previous algorithm benchmark from 0.95225 to 10% for a $1 million prize. With such a small improvement on the existing recommendation system, the company can increase more revenue [2]. Established companies such as OpenAI and GitHub Copilot take recommendation system to another level by developing cloud-based artificial intelligence tool which provides autocomplete-style suggestion in an attempt to make coding easier and more efficient so that the companies can maximize their profits. In this paper, we first describe, explain, and explore a wide range of methods and algorithms used in recommendation systems including Content-Based Filtering, Collaborative Filtering and Matrix Factorization, and Recommendation Using Deep Neural Networks. In contrast to the existing survey, we focus and discuss on the recommendation algorithm that is best fit for podcast applications.**

Keywords— **Recommendation system, Type of the recommendation system, Content-Based Filtering, Collaborative Filtering and Matrix Factorization, Recommendation Using Deep Neural Networks**

1. Introduction

A recommendation system is a software tool and algorithms providing suggestions to users based on their interests to help them find new products and services they might be interested in by filtering out information that is not relevant to their specific tastes or preferences. Nowadays, e-commerce websites have been flourishing extremely fast and allowing people to buy and sell a lot of objects such as physical goods, services, and digital products over the internet. When there are so many things to choose from, it's important to use an extra tool called the Recommendation system. The recommendation is normally based on various decision-making processes, such as purchasing a product or reading the news. The recommendation system gives customers a chance to find things that they might not have found on their own. It keeps track of what each customer likes and then shows those items to the customer. Recommendation systems have become more popular in recent years, and they are now used in movies, news, books, music, search queries, and merchandise in standard. Recommendation systems benefit both provider companies and users. Provider company they don't need market research to find out whether a customer is willing to purchase at a shop where they're getting maximum help in scouting the right product. And users generally like to be suggested things that they would like, and when they use a site that can relate to his/her choices extremely perfectly then he/she is bound to visit that site again [6].

E-commerce sites use recommendation systems to show their customers about products they might be interested in. The products can be recommended based on the top overall sellers on a site, the customer's demographics, or an analysis of the customer's previous buying behavior to predict what they will purchase in the future. In general, these techniques are part of personalization on a site because they help the site adapt to each customer. Recommendation systems automate personalization on the Web, which means that each customer can have their personalization [11].

The research paper is helpful to two groups: researchers looking at recommender systems in e-commerce and site owners who are thinking about implementing them in their websites. All the methods and examples serve as a helpful starting point for academics to situate their research. The framework will surely be enlarged to incorporate the best recommender system application. The paper offers implementers a method for selecting from among the applications and technologies of all methods. Especially, we will propose one method that is compatible with the podcast application, and explain why we choose one of them.

2. Literature Review

Since the earliest days of computing, people have been exploring the idea of using computers to make predictions on what would be most beneficial to the user. The concept of recommendation systems (RS) was initially put into practice for the first time in 1979 in a program known as Grundy, which was a computer-based library that offered recommendations to the user regarding which books they should read [3]. According to the research paper that was released in 1998 and titled "Recommender Systems: A GroupLens Perspective,"[4] In the early 1990s, the GroupLens Research Project began exploring possible approaches to using the perspectives of a group of people to help individuals filter content in a better way. Based on the fields in which they specialized, there were two key research specializations: (1) With Artificial Intelligence (AI), the team researched to develop technologies that would serve as a "knowledge robot," or "knowbot," that genuinely searches out the information, analyses it, and response with the information that the knowbot would decide would be the most

beneficial to its user [4]. (2) Information Filtering (IF), the team conducted research with the goal of developing technologies that would be more effective in identifying articles that contained keywords likely to be of interest to its user. Even though these technologies have produced good results and also have been beneficial to users, they still have one major drawback. In the case of the knowbot, that technology has not yet advanced to a level where it can do a great job of analyzing articles as well as a human does. In the case of information filtering, identifying thousands of articles with keyword-based searches is not a scalable solution to finding articles that contain any imaginable set of keywords. The weaknesses of these two existing filtering systems, taken together, created an opportunity for a new type of filter that would focus on finding a list of articles that match the tastes and interests of each user would like, regardless of their content. The GroupLens research group has decided to apply a new method of filtering information to Usenet news when hundreds of thousands of articles are posted daily [4]. This method would be more effective than existing filtering systems that are currently in use in Usenet news. With the development of the "GroupLens" recommender, which is now known as automatic collaborative filtering, they were among the first to study automated recommender systems [5]. It collects ratings from people who have read an article, calculates the average rating of each article, and recommends the articles most likely to appeal to readers based on their preferences. To design a prediction system to rate articles, they use single-dimensions rating, with the dimension being "What score would you have liked GroupLens to predict for you for this article? They found that users were more likely to read articles that GroupLens predicted they would enjoy, and this tendency was stronger than the tendency to read either randomly selected articles or articles that GroupLens predicted they would dislike. Because users tended to dislike most articles and because GroupLens was effective at identifying articles that users would like, users requested the ability to scan a newsgroup for high-rated articles. This led to their exploration of a different style of interface for collaborative filtering systems: TopN interfaces. Rather than providing a score for each article, TopN interfaces greedily can find a set of articles likely to be highly rated by an individual user [4].

Gupta and Katarya [10] explain that collaborative filtering is a technique in recommendation systems in which the recommendations are dependent on the user's neighbors, and this technique uses matrix factorization to create a matrix containing users, items, and ratings from those items with different kinds of users. Collaborative filtering has been used in many e-commerce platforms and provides a better experience than other techniques. The following diagram illustrates the process:

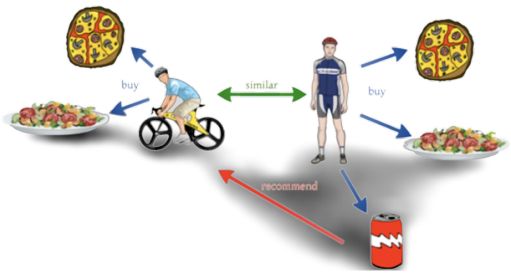


Fig. 1 Collaborative Filtering [12]

There have been two main directions in the evolution of recommendation systems: collaborative filtering and content-based filtering. Collaborative filtering is a method of recommendation that uses algorithms to filter data from users' reviews and make personalized suggestions for users who have similar preferences. Content-based filtering uses a user's browsing history to recommend other items similar to what the user has previously liked, based on their past actions or explicit feedback. For instance, if a user likes to watch movies such as Spider-Men, the recommender system will recommend movies of the superhero genre. Thus, every time a user like another genre of movie, this new information is added to their profile. In a case in which there is known data on an item (such as its name, location, or description), but no information about the user, content-based methods are the most effective solution compared to the collaborative method.

According to Po-Wah Yau and Allan Tomlinson [8], Content-based filtering works by first analyzing the quality of an item, then matching its properties to those in a database. To do this, content-based filtering algorithms rely on keywords to predict which items a user would like, based on the user's ratings in past.

In a content-based filtering survey by Mladenic [9], it was found that the technique works by first searching for items similar to those the user has previously expressed interest in, then constructing a model based on these previous interests. This model generates recommendations for future purchases. The following diagram shows the algorithm's process in e-commerce websites.

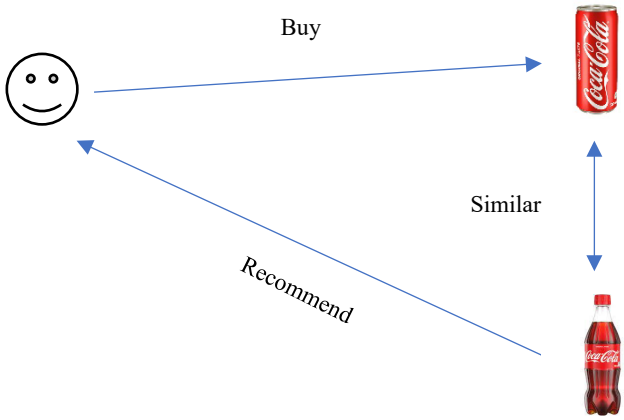


Fig. 2 Content-Based Filtering

3. Recommendation System

In this part, we are going to describe and demonstrate various methods of content-based filtering recommender systems using dot product as a similarity measure, and collaborative filtering (CF), such as user-user, item-item, and rating prediction. Because of the limitation of Existing methods in recommendation systems, we are going to see how Deep learning is used in recommendation systems. The following section will be dedicated to content-based method and how they work.

3.1 Content-Based Filtering

To demonstrate content-based filtering, now let's go through the entire concept of content-based filtering. In

content-based filtering the first thing is the platform tries to understand what a user likes then it tries to derive the type of content that the user is liking, then it looks into its entire directory and chooses the content which is very much relevant to the user, then it finds the best fit of content and then it recommends it to the user. Figure 3 shows an example a scenario of a movie recommendation, the user has watched 3 movies in the romance category and now the user wants to watch the fourth movie, so if the user asks the system for the fourth movie it will look for other movies that the user has watched, then it will look in its directory what are the romance movies it is having, based on that it will make the recommendation for the user. Another way to approach it, this was very vague just because the user has watched a romance movie does not mean that the user has liked it. Another way to approach this problem is by analyzing the rating that a user has given the feedback. suppose the user has watched three movies and for the three movies the user has given the following feedback:

Fig. 3 User rating

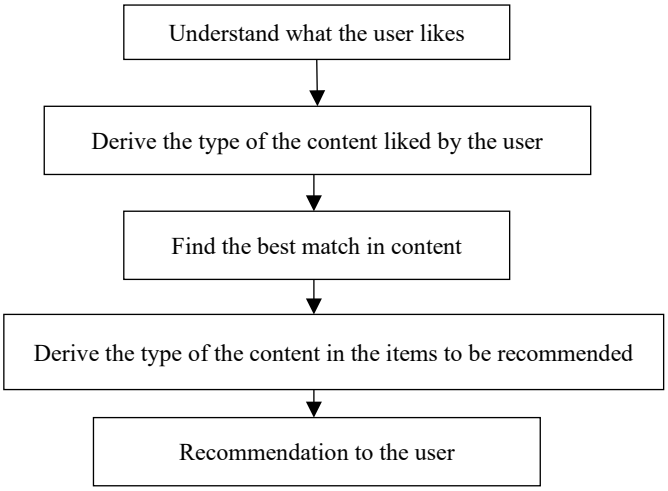
We can see that users give a rating to **A** movie 9/10, **B** movie 4/10, and **C** movie 5/10. Now the system will most likely recommend the use of the type of movie which has got the highest feedback. A good recommendation system should recommend a movie that is related to the type of movie to which the user has given the highest feedback.

A. Using Dot Product as a Similarity Measure

We all are familiar with vectors: they can be 2D, 3D, or whatever-D. let's think in 2D for a moment, because it's easier to picture in our mind, and let's refresh the concept of dot product first. The dot product between two vectors is equal to the projection of one of them on the other. Therefore, the dot product between two identical vectors (i.e., with identical components) is equal to their squared module, while if the two are perpendicular (i.e., they do not share any directions), the dot product is zero. Generally, for n-dimensional vectors, the dot product can be calculated as shown below.

$$u.v = [u_1\ u_2\ \dots\ u_n].\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_3 \end{bmatrix} = u_1v_1 + u_2v_2 + \cdots + u_nv_n = \sum_{i=1}^{n} u_iv_i$$

The dot product is important when defining the similarity, as it is directly connected to it. The definition of similarity between two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ is, in fact, the ratio between their dot product and the product of their magnitudes. To demonstrate content-based filtering now let's break this flowchart into fourth different steps.



1) *Step-1 Understand what the user like*: In Table 1 Suppose we have these three movies that we had named earlier **A**, **B**, and **C** these are the ratings that the user provided. To understand what a user like this is done by looking at the below data set.

TABLE 1
USER RATING INPUT

Movie Name	Rating
The Avengers	8
The Faults in our stars	4
Jurassic Park	6

From our data, we understand that our user likes the movie the avengers but does not like The Fault in our stars.

2) *Step-2 Derive the type of content liked by the user*: The system does not understand the name of the movie, so we try to understand what type of content the user is liking so we break down the content into categories.

TABLE 2
MOVIE

Movie Name	Super Hero	Adventure	Tragedy	Sci-fi	Romance
The Avengers	1	1	0	1	0
The Faults in our stars	0	0	1	0	1
Jurassic Park	0	1	0	1	0

What we do is we have a certain genre for movies we have super hero Adventure, Tragedy Sci-fi and Romance. The following Table 2 shows a feature matrix where each row represents a list of movies and each column represents the type of movie. The type of movies could include categories (such as Super Hero, Adventure, Tragedy, Sci-fi, and Romance) The 1/0 simply represents whether the movie has adventure/action, etc., or not. This type of representation is called a binary representation of attributes.

To detect similarities between movies, we need to vectorize. Initially, we have the user rating matrix then we have the movie matrix. So, the next thing we do we have to multiply row-wise all the values. Doing this way, we can generate a weighted genre matrix as shown in Table 3.

TABLE 3
WEIGHT GENRE

Movie Name	Super Hero	Adventure	Tragedy	Sci-fi	Romance
The Avengers	8	8	0	8	0
The Faults in our stars	0	0	4	0	4
Jurassic Park	0	6	0	6	0

Once we have generated the weight genre matrix, the next thing is to try to understand what type of genre the user likes we can do that by applying the summation of column-wise as shown in Table 4.

TABLE 4
CREATE USER PROFILE

	Super Hero	Adventure	Tragedy	Sci-fi	Romance
User	8	14	4	14	4

After applying the summation, we can see that the user gave 8 for super Hero, 14 for Adventure, 4 for Tragedy, 14 for sci-fi, and 4 for Romance. In Table 5 we can perform normalization to the weighted genre matrix, the reason we do normalization operation is that we try to map the huge dataset to a smaller value. After the normalization operation, we can create a new type of matrix called a user profile matrix. By doing that we have understood what type of content is liked by the user.

TABLE 5
USER PROFILE AFTER NORMALIZATION

	Super Hero	Adventure	Tragedy	Sci-fi	Romance
User	0.057	0.37	0.34	0.37	0.34

3) *Step-3 Derive the type of content in the items to be recommended*: Now that the platform is going to look at its directory and try to find what type of content it is having. In the below figure, In Table 6 we have created a new matrix called the candidate matrix in which we have 4 movies that we have to recommend next to the user.

TABLE 6
THE CANDIDATE MOVIE

Movie Name	Super Hero	Adventure	Tragedy	Sci-fi	Romance
Titanic	0	0	1	0	1
The Lion King	0	1	0	0	0
Spiderman	1	1	0	1	0
Harry Potter	1	1	0	0	0

4) *Step-4 Find the best match in content*: To find the best match we simply again perform a little bit of matrix multiplication by multiplying the user profile matrix in Table 5 with the candidate movie matrix in Table 6. Then in Table 7 we do a column-wise multiplication with the candidate movie matrix that we had generated in step 3.

TABLE 7
WEIGHTED MOVIE

Movie Name	Super Hero	Adventure	Tragedy	Sci-fi	Romance
Titanic	0	0	0.34	0	1
The Lion King	0	0.37	0	0	0
Spiderman	0.057	0.37	0	0.37	0
Harry Potter	0.057	0.37	0	0	0

To generate a weighted movie matrix, we have to take a row-wise summation and by doing that we finally get the recommendation table which we are going to use to make the recommendation to the user as shown in Table 8.

TABLE 8
RECOMMENDATION MOVIE

Movie Name	Rating
Titanic	0.68
The Lion King	0.37
Spiderman	1.31
Harry Potter	0.94

Every candidate movie has got some type of numerical rating and the movie with the highest rating is the movie that the recommended system should recommend to the user.

3.2 COLLABORATIVE FILTERING AND MATRIX FACTORIZATION

In the previous section, we looked at content-based approaches to building recommendation systems. In this section, we are going to look at another popular approach called collaborative filtering. The basic idea behind collaborative filtering is very simple. Suppose we have a user ***X*** whom you want to make a recommendation, what we're going to do is we're going to find a group of other users whose likes and dislikes similar to the user ***X***. For example, suppose we are doing movie recommendations. Now this group of users like the same movie that the user ***X*** like and dislike the same movie that the user ***X*** dislike. We call this set of users **the neighborhood** of user ***X***. Once we find the set ***N*** of users or the neighborhood of the user that is similar to the user ***X***, we find other movies that are liked by a lot of users in the set ***N*** and recommend those items to the user ***X***. So that is the basic idea behind collaborative filtering. The key trick is to find the set of users that are similar to the user ***X*** by defining a notion of similarity between users.

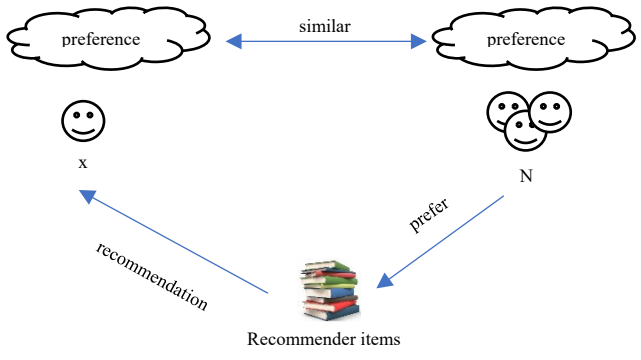


Fig. 4 Collaborative filtering

A. User to User

Here in Table 9 is a simple example with 4 users **A**, **B**, **C**, and **D** with a bunch of movies.

TABLE 9
MOVIE

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

For example, user A has rated 3 movies **HP1**, **TW**, and **SW1**. Imagine they are rating on a scale from 0-5 stars. Users **B** and **C** have rated 3 movies, and user **D** has rated 2 movies. Let's call the set of ratings for a user "The user rating vector". Consider we have 2 users $\boldsymbol{X}$ and $\boldsymbol{Y}$ with rating vectors $\boldsymbol{r_x}$ and $\boldsymbol{r_y}$**.** What we need is a similarity matrix $\boldsymbol{sim(X,Y)}$**.** Now the interesting here there are a lot of movies that user **A** has not rated and lots of movies that **B** has not rated. So, the key to defining the similarity function is how we deal with unknown values. Once we define the rating vector, we'd like to define it in a way such that it captures a very simple intuition that users with similar tastes have higher similarity with users with disabilities. For example, in the case where users **A** and **B** have rated only 1 movie in common, however, they both rated the movie fairly highly whereas users **A** and **C** have rated 2 movies in common, but their ratings are very dissimilar which user **A** seems to like **TW** while user **C** doesn't like it while user **A** hate **SW1** and user **C** loves it. It seems intuitive that users **A** and **C** are dissimilar while users **A** and **B** are similar. So, we need to capture this intuition when we define the notion of the similarity of **A** and **B** to be higher than the similarity of **A** and **C**. To do this there are 3 options:

1) *Option-1 Jaccard Similarity*: A The first option to try by using Jaccard similarity:

$$\boldsymbol{sim(A,B)} = \frac{|\boldsymbol{r_A} \cap \boldsymbol{r_B}|}{|\boldsymbol{r_A} \cup \boldsymbol{r_B}|}$$

Here we just take the intersection of rating vectors and divide it by the union. Notices when using the Jaccard similarity A and B have 1 rating in common (HP1) of all 5 movies and therefore the

$$\boldsymbol{sim(A,B)} = \frac{\boldsymbol{1}}{\boldsymbol{5}}; \boldsymbol{sim(A,C)} = \frac{\boldsymbol{2}}{\boldsymbol{4}}$$

Notice though that when we calculate similarities in this way the Jaccard similarities of **A** and **B** are less than the Jaccard similarities of **A**, **C**. $\boldsymbol{sim(A,B) < sim(A,C)}$**.** This is counter to the intuition that we wanted we want to capture that **A** and **B** are more similar than **A** and **C**. So, we'll have the abandon this notion of Jaccard similarities. As we can see the problem with the Jaccard similarity that we like to fix is that it ignores the rating values, it just notices that **A** and **B** have not watched one movie in common while **A** and **C** watched 2 movies in common, but it doesn't know that fact that how the user like the movies that they watched.

2) *Option-2 Cosine Similarity*: A way to capture the rating values and use them to compute similarity by using **Cosine**, so that we can compute the cosine between the vectors.

Let's say we define the $\boldsymbol{sim(A,B) = cos(r_A, r_B)}$**.** To compute the cosine similarity, we have to insert some value for the unknown ratings, and the easiest way to do this is to insert them as **0** as shown in Table 10**.**

TABLE 10
MOVIE

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4	0	0	5	1	0	0
B	5	5	4	0	0	0	0
C				2	4	5	
D		3					3

In this case the $\boldsymbol{sim(A,B) = 0.38, sim(A,C) = 0.32}$**.** As we can see the similarity between **A** and **B** is greater than the similarity of **A** and **C** as we want it, but not by much.

The problem that we have with Cosine Similarity is that it treats the missing ratings as a negative rating. What we've done is that we've used 0 to fill in the blank. On our rating scale from 0 to 5, 0 is the worst possible rating. Consider if user **A** has rated **HP2** for 0 which is a bad assumption given the fact that they like **HP1**. This is the problem that we have with the Cosine rating.

3) *Option-3 Centered Cosine:* One way to fix Cosine similarity to accomplish what we want is to use Centered cosine. In Table 11 to do that we're going to normalize the rating for a given user by subtracting the row mean of the average rating of the user and treating the blank rating as 0.

TABLE 11
MOVIE

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	2/3			5/3	-7/3		
B	1/3	1/3	-2/3				
C				-5/3	1/3	4/3	
D		0					0

If we try to sum up the rating in any row, we'll be getting 0. What we've done is that we've centered the ratings of each user around zeros. So, zeros become the average rating for every user, and positive ratings indicate that the user liked a movie more than average and negative ratings indicate that the user liked the movie less than average the magnitude of the rating are also showing how much the users liked or disliked a specific item.

Once we did this centering, we can begin to compute cosines using these centered ratings. After we do that the $\boldsymbol{sim(A,B) = cos(r_A, r_B) = 0.09; sim(A,C) = -0.56}$ which captures the fact that **A** and **C** are quite dissimilar users. It means that the movie that user **A** like user **C** doesn't like and the movie that user **A** doesn't like **C** likes. As we can see now there is a big gap between the similarity of **A** and **B** and the similarity of **A** and **C**. It shows that users **A** and **B** are much more alike than users **A** and **C**.

So, the centered cosine captures the intuition of similar users much better than the simple cosine, because the missing ratings instead of being treated as the negative rating are treated as

average ratings. It also turns out to be a nice way to handle tough raters and easy raters because some data tend to be tough raters by giving a great movie on a scale of 0 to 3 while others tend to be easy raters that tend to be much more liberal with a star rating. By subtracting out the average rating of each user we've centered users around an average of zero and we normalized some of the tough raters and the easy raters.
As we can see the centered cosine has these two advantages of centering around zero and capturing our intuition better. Centered Cosine in the statistic world is also known as **Pearson Correlation**.

Rating Predictions

We've come up with a way of estimating similarity between users but how do we make rating predictions for a user?
Let r_x be the vector of user X ratings. We're going to use the notion of centered cosine similarity to find the set N of users which we called the neighborhood. The neighborhood contains the K users who are more similar to X users. We are going to go through the set of all users to computer the similarity between user X and select the top K users with the highest similarity value that we called the set N. But to be careful since we are trying to estimate the rating of item i by the user X. We have to make sure that the set of N contains only users who've rated item i. Once we have this, we can make a prediction for user X and item i.

1) *Option 1:* The simplest prediction is to take the average rating from the neighborhood. As we know the set N contains users who also rated item i and are similar to user X. The simplest estimate is to just use the average rating of all the users for item i in the neighborhood then take that as our estimate of the rating for user X and item i.

$$r_{xi} = \frac{1}{k \sum_{y \in N} r_{yi}}$$

2) *Option 2:* Now option 1 is very simple but it ignores the actual similarity values between users while neighborhood N contains users who are similar to item i. There might be a range of similarity values within the neighborhood that might contain users who are very highly similar to the user X and a few users who are not that similar to the user X. now what we like to do is to weight the average rating by the similarity values and that gives us to option 2:

$$r_{xi} = \frac{\sum_{y \in N} s_{xy} r_{yi}}{\sum_{y \in N} s_{xy}}; \; s_{xy} = sim(x, y)$$

Option 2 is a weighted average. We can see the neighborhood N and for each user Y in the neighborhood N we weighted wise rating for i by the similarity of X and $X + Y$, then we just normalize it by the sum of the similarity. Doing that gives us a rating estimate for user X and item i .

B. Item to Item

Another technique in collaborative filtering is item-to-item. The basic idea is simple, instead of starting with the user and finding similar users we're going to start with an item i and find other similar items to the item i then we estimate the rating for item i based on the ratings for the similar items. We can use the same similarity metrics and prediction functions as in the user-user model.

Here is the rating function:

$$r_{xi} = \frac{\sum_{j \in N(i;x)} s_{ij} r_{xj}}{\sum_{j \in N(i;x)} s_{ij}}$$

What it trying to do is to predict the rating for user x and item i. It starts with item i and we're going to find a neighborhood of items. The neighborhood of items $N(i; x)$ is a set of items that are both rated by the user x and similar to the items i that we're looking at. In that way, we just take the item i to compute the similarity between item i and every item in the set that we are known about restricting the attention only to the items that have been rated by the user x and then take the top k of those as a neighborhood $N(i; x)$. Once we do that, we can just the same weighted average formula that we had in the user-user approach to predict the rating of user x and item i.

Example: item-item CF (\|N\|=2)

Let's say the neighborhood size that we want to pick is the size of n at the 2 nearest neighbors of item i.

Table 12 is the utility matrix. The yellow color is the known rating in the utility matrix and the white color is the unknown rating. We have movies on the y axis and numbers of users on x axis.

TABLE 12
MOVIE (Y-AXIS), THE USER (X-AXIS)

	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3			5			5		4	
2			5	4			4			2	1	2
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			3	
5		4	3	4	3						2	5
6	1		3		3			2			4	

Unknown rating Rating between 1 to 5

Table 13 Let's assume the rating value is between 1 to 5 and our goal is to estimate the rating of movie 1 by user 5.

TABLE 13
MOVIE (Y-AXIS), THE USER (X-AXIS)

	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3		?	5			5		4	
2			5	4			4			2	1	2
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			3	
5		4	3	4	3						2	5
6	1		3		3			2			4	

Estimate rating of movie 1 by user 5

The first step is to take movie 1 and find another movie that is similar to movie 1. To do that we are going to use the **Pearson correlation** as the similarity which is the same as a centered

cosine. We are going to take every other movie and compute its centered cosine from movie 1. And we are doing that in movie 1 the centered cosine is 1.00, movie 2 is somewhat dissimilar to movie 1, and so on.

TABLE 14
MOVIE SIMILARITY

Movies	sim (1, m)
1	1.0
2	-0.18
3	0.41
4	-0.10
5	-0.31
6	0.59

In Table 14 since our neighborhood's size is 2 we need to find the 2 movies with the highest similarity to movie 1 and also been rated by user 5. And those happen to be movie 3 with a similarity of 0.41 and movie 6 with a similarity of 0.59. So, we're going to pick those 2 movies as a neighborhood for movie 1. Once we do that the similarity value is

$$S_{13} = 0.41;\ S_{16} = 0.59$$

we're going to use a weighted average estimate to predict the rating for movie 1 and user 5.

The weighted average:

$$r_{15} = (0.41 * 2 + 0.59 * 3)/(0.41 + 0.59) = 2.6$$

TABLE 15
MOVIE (Y-AXIS), THE USER (X-AXIS)

	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3		2.6	5			5		4	
2			5	4			4			2	1	2
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			3	
5		4	3	4	3						2	5
6	1		3		3			2			4	

Finally Table 15, 2.6 is the predicted rating for movie 1 with user 5.

4 PROPOSED WORK

Podcasting has established itself as one of the most powerful tools for sharing new ideas in the digital age. With over half of the US population as its listener base, podcasting has taken the internet by storm. This is due in no small part to the fact that a podcast can be almost anything the creator wants it to be. Unlike traditional radio, there are no rigid guidelines to follow in terms of air-time, advertising, or even release schedules. This has created a cultural environment in which the creators can take center stage and focus on high quality truly unique content. A podcast is a form of on-demand, internet talk radio that is focused on a particular topic and is delivered in an episode format.

We all know about podcasts that are specified in some of the popular apps like Spotify, apple podcasts, google podcasts, etc. We also know that recommendation means suggesting to others based on the information we acquired that might be useful for others. In addition, we all know there are many types of recommendation systems such as the traditional approach and the Modern approach, However, what remains unknown is which recommender system method will be best for different types of platforms or businesses, particularly for podcasts.

For the podcast application, we think that Content-Based Filtering is the best method that is entirely compatible with it. There are 2 major benefits of using content-based filtering in part of the podcast application such as personalization for new users and shuffle mode.

- *Personalization for the new user*: one of the most obvious ways to do this is to force new users when they first sign up to make a few selections by choosing their favorite creators and genres to get started in the application. that is the way that we can get the user's interest, and then the system will provide recommendations with high accuracy based on the user's selection. Content-Based Filtering does not need any data about other users, since the recommendations are specific to this user. This makes it easier to scale to a large number of users.
- *Shuffle mode*: In shuffle mode when the application started, it operated on true random shuffle. Every Podcast playlist had an equal chance of coming up when you pressed the shuffle button. However, the shuffle wasn't random enough, and they kept getting clumps of podcasts by the same creator when on shuffle. To combat this, we can change the shuffle to be less random by selecting content-based filtering as an algorithm. To do this we select first a random playlist and analyses each podcast in the playlist and then match podcast which has the similarity attributes, such as genre or tags. If a listener like a particular podcast, then the content-based system will recommend podcasts that are similar to that podcast.

There are several ways to build a Recommendation and these approaches are divided according to the need of the application. Though there are several types of Recommendation Systems, Content-Based Filtering is still the best method for Personalization for the new user and Shuffle mode. There are 3 reasons that make Content-Based Filtering the best choice:

1) Content-based recommendation systems provide user independence through exclusive ratings which are used by the active user to build their profile.
2) Content-based recommendation systems provide Transparency to their active user by giving an explanation of how the recommender system works.
3) Content-based recommendation systems are adequate to recommend items not yet placed by any user. This will be advantageous for new users.

5 CONCLUSION AND FUTURE WORK

In this paper, we try to describe the various type of recommendation techniques and its type with examples. We also proposed compatible techniques for the podcast application system.

In the future, we can develop and assess a variety of other attributes and techniques to facilitate the effective implementation of recommendation systems. Additionally, by

merging recommendation systems with machine learning (ML) and natural language processing (NLP), we will construct powerful and efficient recommendation systems that rely on the experiences it's had in the past. This will result in an intelligent recommendation system that anticipates what is in the user's best interest and, as a consequence, provides recommendations with a high level of accuracy.

For our future work, we will implement the content-based filtering algorithm that we have been proposed above for the podcast application.

ACKNOWLEDGMENT

We are extremely grateful to our advisor Mr. Kor Sokchea for his invaluable patience and feedback. And we also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

We are also grateful to our team members, for their peer-review our paper, late-night feedback sessions, and inspiration.

REFERENCES

[1] Das, D., Sahoo, L., & Datta, S. (2017). *A survey on recommendation system.* International Journal of Computer Applications, 160(7).

[2] Carlos A. Gomez-Uribe and Neil Hunt. 2016. *The Netflix Recommender System: Algorithms, Business Value, and Innovation.* ACM Trans. Manage. Inf. Syst. 6, 4, Article 13 (January 2016), 19 pages.

[3] Qomariyah, N. N. (2018). *Pairwise Preferences Learning for Recommender Systems* (Doctoral dissertation, University of York).

[4] Konstan, J. A., Riedl, J., Borchers, A., & Herlocker, J. L. (1998). *Recommender systems: A grouplens perspective. In Recommender Systems: Papers from the 1998 Workshop* (AAAI Technical Report WS-98-08) (pp. 60-64). Menlo Park: AAAI Press.

[5] Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. (1997). *Grouplens: Applying collaborative filtering to usenet news. Communications of the ACM*, 40(3), 77-87.

[6] S. C. Nair, "*A Comprehensive Survey on Recommendation Systems Based on Collaborative Filtering*," 2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS), 2022, pp. 1-5, ~~doi~~Doi: 10.1109/IC3SIS54991.2022.9885627.

[7] Ricci, F., Rokach, L., Shapira, B. (2022). *Recommender Systems: Techniques, Applications, and Challenges.* In: Ricci, F., Rokach, L., Shapira, B. (eds) Recommender Systems Handbook. Springer, New York.

[8] [7] P. W. Yau and A. Tomlinson, "*Towards Privacy in a Context Aware Social Network Based Recommendation System*," Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, Boston, MA,2011, pp.862-865. Doi:10.1109/PASSAT/SocialCom.2011.87

[9] Mladenic, D.: Text-learning and Related Intelligent Agents: *A Survey. IEEE Intelligent Systems* 14(4), 44–54 (1999)

[10] *Recommendation analysis on Item-based and Userbased Collaboration Filtering* Garima Gupta, Rahul Katarya, India

[11] Ben Schafer, Konstan, & Riedl. (n.d.*). Recommender Systems in E-Commerce. ACM Digital Library*. Retrieved November 16, 2022.

[12] Kirk, J. (2019, January 22). *Getting started with recommender systems and TensorRec. Medium*. Retrieved December 20, 2022.

Design Microservices Architecture for Podcast Application

Pich lyheang[#1], Kor Sokchea[*2]

[#]*Department of Information Technology Engineering, FE, Royal University of Phnom Penh*
Phnom Penh, Cambodia
[1]2820.pich.lyheang@rupp.edu.kh

[*]*Corresponding author: Kor Sokchea*
[2]kor.sokchea@rupp.edu.kh

***Abstract*— Software architecture emerging as subdisciplines of software engineering is defined as the fundamental structure of a software system consisting of software elements, the externally visible properties of those elements, and the relationships among them [1][2]. With a good software architecture, it enables software developers to develop a software system that provides best performance, scalability, and availability as well as without code redundancy. In this paper, we propose the design of system architecture for our Podcast application called Akara by using one of the most popular architectural patterns called microservices. In addition to technical detail, we will discuss about reasons we opt microservices for our system.**

***Keywords*— Software Architecture, Related work, Architectural Style, Microservice Architecture, Podcast, Akara.**

I. Introduction

Software architecture, known as subdiscipline of software engineering, is a about studying of how software systems are designed and bult. It has been seen in high demand job in the software industry. [12] Because of rapid evolution of technologies, the people around the world need to transform the way they live and the way they work by using and assigning all the stuffs to be controlled and worked. With the help of technologies, a wide range of companies and organizations have shifted from traditional workflow by integrating technology in operation. Every company or organization basically has their own strategic planning of success which strongly requires software development has to follow business requirements in effort to achieve the company or organization's goals. In order to respond to global demands for high quality, complex, and large software, software industry also demands their software developer to possess a wide knowledge of software development life cycle as well as software architecture skill. With this reason, software architecture plays a vital role to build a vast system successfully and efficiently. Moreover, the complexity of the business is inevitably happened as well as constantly change to quickly achieve their goal, so the businesses certainly need a well-thought and flexible software architecture satisfying the business requirements to help them avoid any anticipated issues. That is why, the existing of many architectural styles exist and align with a specific purpose of a particular business. In earlier decade, there is a popular and powerful architectural style appeared to reduce lots of business' concerns as well as be able to apply in several conditions of business called microservice architecture. Therefore, in this work, we propose the use of microservices architecture to design system architecture for our podcast application called Akara.

II. Related work

To enhance the process of comprehension of this paper, we would like to bring your attention to realize some considerable background related to how some enormous companies have done and tackled with their built software's decision making to align with their flexible business change and to see how amazing those businesses obtained by conforming the architectures they applied. All of those companies you certainly recognize such as Amazon, Netflix, Uber, Spotify and more. These companies applied microservice architecture in their system; thus, in this paper, we only chose one of them to illustrate that is Amazon company. Why this enormous online e-commerce giant chose microservices architecture to be applied in their system?

All of you mostly know amazon is an internet retail giant, but it did not start that way. Originally, Amazon commenced with monolithic architecture, every component is tightly interdependent. During doing this architecture, whenever Amazon needs to upgrade or scale the system, they faced challenging with lots of excruciating condition because of its entangled dependencies. As a result, it causes the company to put more effort, money and time into solving those issue. In 2001, the company was facing with a severe circumstance with their growing of customers using the system causing the system necessarily need to be scaled. Intelligently, Amazon broke its monolithic application into small components running solely and service-specific based. That moment was the occurrence of microservice. The way of amazon's developers came to be rather active by analysing, working out and determining which sections could be turned into microservices. They separated sections of code and wrapped these units in a web service interface. Amazon assigned ownership of each independent service to a team of developers allowing them to view development bottlenecks more granularly. By doing so, it made all the team developers could solve challenges more efficiently since a small number of developers could direct all of their attention to single service. Based on this example, you can see that microservices come and enable Amazon to be able to deal with many struggling stuffs it faced [9]. Without microservices, Amazon could not have grown to be the most valuable company in the world valued market worth 1.6 trillion dollars of August 1,2022. Consequently, according to the experience of being applied microservices in Amazon, it is literally obvious that

microservice accompanies with a wide range of benefit especially when it is used to deal with complex system.

III. Software Architecture

The software architecture of a system or a collection of systems consists of all the important design decisions about the software components and the interactions between those components that exist in the systems [6]. Software architecture can be used as a foundation to make strategic decision of the created software with business strategic plan. Furthermore, it also provides a principal constraint of the software need to be complied by. Software architecture aids software development to prevent any unanticipated problems during implementation and after system completion.

There are some of essential principles which are given from implementing the architecture [7]:

- Separating concerns
- Encapsulation
- Single responsibility
- Dependency inversion

A. Architectural Style

Architectural styles and patterns are defined the way how to organize the components of the system so that one can build a complete system and achieve the requirements of the customer [5]. Regard to what mentioned earlier in the software architecture section, there are lots of styles of architectures used in software industry based on the requirement domain of the business. Choosing the architecture style is an absolutely crucial point should be conscientiously considered when making decision related to which architecture should be selected to handle the whole software infrastructure because opting an inappropriate architecture can negatively impact to the entire software application. Different architecture has different advantages, so please take a look at the below list showed about the popular architectural style have been used in development of software design.

Common architectural styles are frequently utilized [5]***:***

- Monolithic architecture
- Service-oriented architecture
- Serverless architecture
- Client-server architecture
- Layer-based architecture
- Event-driven architecture
- Microservice architecture

1) Microservices architecture:

Microservices architecture is an approach in which a single application is composed of many loosely coupled and independently deployable smaller services [3][4], and it is also a style of engineering highly automated, evolvable software systems made up of capability-aligned microservices [4]. Microservices are likely to be at least as popular with executives and project leaders as with developers. This is one of the more unusual characteristics of microservices because architectural enthusiasm is typically reserved for software development teams. The reason for this is that microservices better reflect the way many business leaders want to structure and run their teams and development processes. According to the survey of over 1200 developers by IBM 87% of the users utilizing the microservice fully agree that the microservice adaptation is worth the expense and effort [3].

There are five core components of microservice architecture should be considered [10]:

- Services
- APIs
- Databases
- Scheduler
- Monitoring

Services are heart of microservices which your logic live, and every service is small, self-contained and perform a specific task. APIs provide communicating way of microservices to clients or others services and exchange data and functionality. Database is where your business's data storages are located, and each service has its own service database specifically. This prevents services from interfering with each other's data. Scheduler roles to manage the running and interacting among services and allow asynchronous running of the services to usefully improve performance. Monitoring ensures that your running services perform properly and collect data to analyse. Monitoring also need to tread all of individual service health to overall system performance. Another importance is that microservices also provides a number of its design patterns to be corresponding to the applied business requirement.

Commonly used pattern of microservices architecture:

- Database per microservice
- Event Sourcing
- Back-end for Front-end (BFF)
- API Gateway
- Saga
- CQRS and etc.

The most significant advantages of microservices architecture [4]*:*

- Scalability
- Availability
- Maintainability
- Failure detection
- Independent deployment
- Composability
- Agility
- Resiliency

IV. Akara Podcast

Akara podcast is an online audio podcast platform is created by a developer team aiming to provide a variety of categories of entertainment for people who are font of audio streaming, refer to any kind of audios. As you can see, the podcast is more modern, flexible, convenient and available than broadcast radio, which used to be well-known, we would always use it in earlier decade, but we have never seen people to listen podcast by the radio as before anymore. Instead, people use podcast application. Every year the population of people come to use podcast application is becoming more and more, and they all need a more modern software and better podcast application to fill their interest. The Akara podcast, therefore, come in.

A. Functionalities of Akara Podcast

In this Akara application consists of five major features which are remarkably noticed, listed below.

- ***Discover feature***: this is a main dashboard of the application to the user when visiting the application, and it contains several sub-sections such as broadcaster section, popular podcast section, history podcast section, gamming podcast section, comedy podcast section and many types more.
- ***Trending feature***: this feature provides a functionality to see the most popular podcasts and podcaster from the system analyse.
- ***Favourite feature***: this feature classifies and store of any users' satisfied podcasts, and it is a pretty convenient way to user to listen what they desire over again.
- ***Playlist feature***: this feature seems to be the podcast organizer due to the fact that this feature collapses all your individual podcast into respectively collective albums to boost the easiness for any users to quickly find the podcast they would like to listen in their favourable podcaster.
- ***Authentication feature***: this feature is basically need when users want to use playlist, favourite and trending features because of the fact that the system requires users to register first before having access to control all of those features.

B. System architecture of Akara

In our goal of designing this software is to find the best architecture fulfilled our system requirement in term of scalability, flexibility, availability and provide an effortless development and deployment when our business grows or completes. Thus, we are eager to walk you through to software design architecture of our Akara podcast now, and let you know how each component of the Akara podcast software interacts from each other. Based on spending much time with finding the architecture, in the end, we chose microservice architecture for podcast application because it is unmistakably adaptable to almost all our need for Akara. This architecture is one of the most suitable architectures among others. You can perceive about how system works and microservice of our Akara application in **figure 1**. For Akara podcast system, we designed all of essential components in Akara podcast into individual layers which potentially help readers effortlessly comprehend and conveniently distinguish how each component of the system processes under the hood. Without a doubt, as you are able to savvy from the below shown figure, there are five layers components are classified such as client layer component, intermediary layer component, microservice layer component, third-party layer component and database layer component. Alternatively, you can design in different way because it differs in different business based on what your business requirements are.

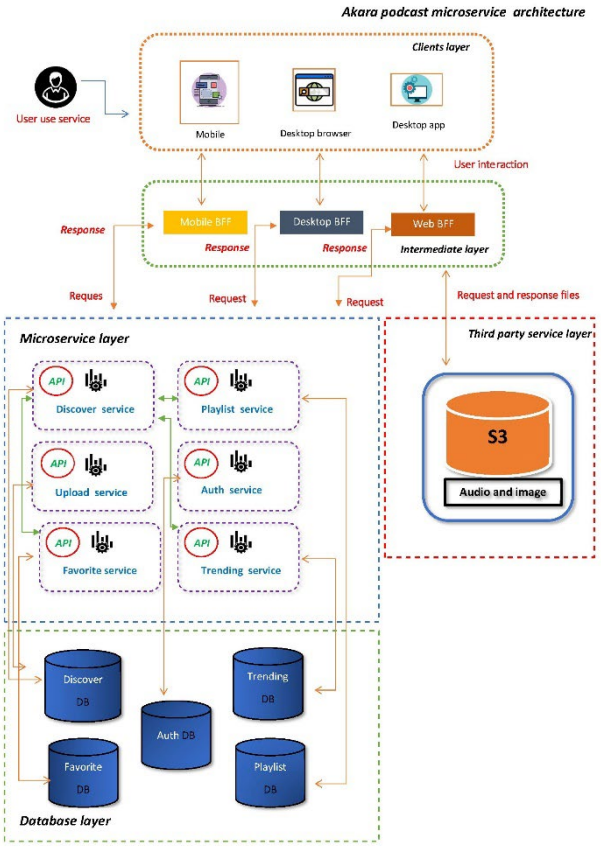


Figure 1 System architecture of Akara

Clients layer section: This section delineated all clients which belongs to our systems. There are three clients of Akarat application. All of those are web-browser based client, mobile-native client, desktop-browser based client. It is followed two approaches to complete successful interface. For the foremost step, each client has its own task to make the request to the back-end side in order to get the data in order to display all those information content for their user interfaces through application programming interface, and they then need to make another request using their data reference getting from the the first mentioned request from the database to make another request so as to get the audio payload from the third-party, S3, it is where every uploaded podcast audio located and database only stored its URL references. Data which each client requires is not the same particularly for mobile-native application, so mobile need to has its own API design for its requirement. by the way, you may curiously ask why the mobile is a bit different from others, web application and desk top app. The answer for this question is that on account of small screen and limited data representation for user interface, undoubtedly, it needs a unique design API for its own. You will know clearer of the separated design of APIs for their user interfaces in the following section. Moreover, when the first request of any services completed, the

next request of the clients just retrieve from the caching located in BFF, you will understand it when you read to intermediate layer. Last but not least, one important party of our podcast client is where we upload data, generally this platform is where admit or podcasters to do all stuff with their podcast uploading. We design in web platform for this special purpose.

Intermediary layer section: For this section, you definitely notice about the design pattern was chosen and applied to microservice-based architecture in Akara and why this pattern is the most appropriate for this business domain. There are three design-patterns which are the most nutritious ones of microservice provided such as direct pattern, API get way pattern, and backend for frontend pattern which commonly known as BFF, all of these are applied to be reverse proxy layer. However, I only walk you through the BFF pattern since it was thought to be the most suitable pattern for Akara podcast application. Perhaps, you all certainly question why it became the best for Akara, so I need you to dive into understanding it slightly. BFF pattern is the next version of API get way pattern. The presence of this pattern exists to furnish extensive loophole missing of API get way pattern. BFF provides a fast development and flexibility to deal with different domains, having multiple clients, and this is a horrible problem and concern within API get way. The primary purpose of BFF is to flexibly design APIs for different clients in virtue of the distinct requirement of data to be display for specific front-end. This pattern, Furthermore, is tightly entangled with front-end, and it mainly aims to provide a better way to front-end with data representing such as preparing and reformatting data for the front-end, and sending the formatted data to the frond-end. Basically, this section is also handled by front-end team to manage the development of BFF. This way makes the development faster because they don't have to wait until the back-end team finishes their task, referring to API development, in order to test the data for front-end. Notwithstanding its advantages, it still has a few disadvantages owing to its latency of multiple requesting from services, but it is not a such concerning issue. Due to the fact that it is much associated with the front-end, it is also deployed altogether. As a result, The BFF is extremely efficient if your system revolves around with more than one client.in contrast, it is not recommended for exclusively one client application. As you know by endlessly mentioned from above of the BFF, whereby being the best for our Akara application. This approach of this pattern extremely meet with our Akara podcast system needs since three distinct clients. Specifically, you can see there are three divergent BFF namely Web BFF, mobile BFF and Desktop BFF. Every BFF is particularly designed to satisfy with their front-end experience. We no long concern about the different requests from the front-end to diverse downstream service since this problem is the reason of BFF comes in and it aggregates all of those requests from miscellaneous API services to a sole API service independently. Thus, when the clients necessitate the data, they just only request to solitary API is sufficient meaning that individual BFF retrieves its own needed data from any microservices and prepare all of those data to be sent to respective user interfaces, Akara web client, Akara mobile client and Akara desktop client, so each client need to request only a single request to BFF, all of those will be stored inside BFF catching when the next time need the data just only get from BFF catching without being tedious to request from database over again. For our Akara mobile client is very convenient to deal with API as you clearly know by the merit of BFF mentioned previously. Moreover, why Akara was not chosen API get ways because API get way still exist the problem of single failure which frequently occurs in monolithic architecture, and it does not suit with multiple client applications [11].

Microservices layer: This section address about individual components of the application, Akara podcast, are organized and divided into services. There are five components of this application need to be concentrated on, all of those components are segregated to distinct services respectively. All of those services namely discover service, favourite service, trending service, playlist service authentication service and backend upload service. Each service roles as a bridge between client forwarding requesting and database processing responds as well as the implementation logic of the entire application live in. All of these services are a RESTFUL API exposing endpoints to the public for the clients or any third-parties are able to retrieve or use the data of Akara. In this case, each endpoint is only referred to the BFF access permission, having been mentioned from above section. The services themselves take on crucial tasks such as authentication, authorization, reverse proxy and catching etc. more significantly, every service is dedicated to one database server meaning that five service use five database server separately making the system more scalable and available. For instance, even though either discover service or playlist, trending service of the system is failed accidentally by technical problem, others services are still alive and serve the rest of services for the user as normal except the problematical service.

Database management layer: There are five databases used in Akara podcast such as discover database, trending database, favourite service, playlist service and authentication service. Discover database is stored the audio data reference, Playlist database is stored the audio playlists data reference, Favourite database is stored the user' interest audios reference data, Trending database is stored the most popular audio data reference of the system and Auth database is used to store the user information. As you can observe from the figure depicted above, this section you are aware of what database pattern is used in Akara podcast. It consists of a variety of database pattern' styles provided by microservice. We, however, use a highly recommended pattern and be applied to our system is called database per service pattern. this is the easiest approach for implementing microservices-based systems, and is often used to migrate existing monoliths with existing databases [8]. This pattern each service belongs to only one database server, so if the problem is occurred by a particular database server, downtime or any issue, it doesn't lead to different services be destroyed simultaneously. Moreover, Doing so is a way to make the Akara podcast loosely coupled and non-dependencies. Particularly, we are fully free to choose any database technologies to apply for each database service when following database per service pattern.

Third-party layer: The section focuses on various services using in Akara podcast. For this case, as you can see, we need only one third-party service is called S3, a popular storage service one of AWS, as our files storage of the entire podcast' audios sound files. Our Akara might need more than one third-party service to assist our system work perfectly. When Akara podcast users upload their podcasts audio from the client those

files will be transferred to be stored in the S3 bucket. hence, all the files audio needed by all the clients are persisted in this storage service. The S3 bucket provides an end-try point for any client to pull files to present in the user interface. The process of deriving the file is done after the process to databases requesting because clients need the data reference first before being able to retrieve all the audio file from the S3 bucket storage.

To summarize the entire process is that when users come to visit our clients, desktop app web app or mobile app, each user interface first will derive the reference audio data from discover database and then the user interface make another request to the third-party where audios are located. After third-party process requests complete, it sends back with the audio source reference for the user interface to display audio for their user, but if users like with any podcasts, and they want to save them, the system will prompt you to register if users have not registered before, Authentication service and BFF will perform this stuff.

V. Future of the application

What we have wholly discussed in this paper is about adopting and designing an architecture for our Akara podcast system, so the next phase is that we are going to implement this proposed architecture within our Akara podcast application system as such.

VI. Conclusion

In summary, choosing the architecture style is what all developers should be though the earliest. There are a number of popular and used architectural styles available, but what you have to keep in mind is that none all of the businesses can be applied those architectures. Opting wrong architecture leads awful impact on your business process. In fact, the architecture applied to Akara podcast application, it is an ideal for this business domain, and its provided quality attributes also meet with the Akara business requirement needed extremely to get a successful system. Therefore, it is undeniable for selecting microservice architecture for Akara podcast. By the way, we also suggest using microservice architecture as Akara podcast does for any podcast application or different complex system. As what we endlessly described almost throughout this paper, you definitely and unavoidably encounter with all of these principles for your built software all of those refer to scalability, availability, flexibility, debug detection and manageability. Consequently, when we need to construct the software, don't forget giving the architectural decision precedence, and if the software is huge and has different sub-system which is arduous to be managed and maintained, the microservice is the best for facilitating your frustrated difficulty.

Acknowledgements

We are extremely grateful to our advisor Mr. Kor Sokchea for his invaluable patience and feedback. And we also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation, and We are also grateful to our team members, for their editing help, late-night feedback sessions. Thanks, research assistants, and study participants from the university, Akara-Podcast team, who impacted and inspired us.

REFERENCES

[1] Bass, L., Clements, P., & Kazman, R. (2003). Software architecture in practice. Addison-Wesley Professional.

[2] Gorton, I. (2011). Essential Software Architecture (No. PNNL-SA-82591). Pacific Northwest National Lab.(PNNL), Richland, WA(United States).

[3] Education, I. C. (2021, March 30). microservices. What Are Microservices? | IBM. Retrieved November 27, 2022.

[4] Alshuqayran, N., Ali, N., & Evans, R. (2016, November). A systematic mapping study in microservice architecture. In 2016 IEEE 9th International Conference on Service-Oriented Computing and Applications (SOCA) (pp. 44-51). IEEE.

[5] Sharma, A., Kumar, M., & Agarwal, S. (2015). A complete survey onsoftware architectural styles and patterns. Procedia Computer Science, 70, 16-28.

[6] Hofmeister, C., Nord, R., & Soni, D. (2000). Applied software architecture. Addison-Wesley Professional

[7] Lindstrom, A. (2006, January). On the syntax and semantics of architectural principles. In Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06) (Vol. 8, pp. 178b-178b). IEEE.

[8] Munonye, K., & Martinek, P. (2020, June). Evaluation of Data Storage Patterns in Microservices Archicture. In 2020 IEEE 15th International Conference of System of Systems Engineering (SoSE) (pp. 373-380). IEEE.

[9] Killalea, T. (2020). A Second Conversation with Werner Vogels: The Amazon CTO sits with Tom Killalea to discuss designing for evolution at scale. Queue, 18(5), 67-92

[10] Dai, W., Wang, P., Sun, W., Wu, X., Zhang, H., Vyatkin, V., & Yang, G. (2019). Semantic integration of plug-and-play software components for industrial edges based on microservices. IEEE Access, 7, 125882-125892

[11] Akbulut, A., & Perros, H. G. (2019, June). Software versioning with microservices through the API gateway design pattern. In 2019 9th International Conference on Advanced Computer Information Technologies (ACIT) (pp. 289-292). IEEE.

[12] Gorton, I. (2006). Essential software architecture. Springer Science & Business Media.

Credit Scoring Model for Cambodian Retail Banking Markets

Meily Oeng, Chanpiseth Chap, Sovila Srun ,Chamroeun Khim

Information Technology Engineering Department, Royal University of Phnom Penh
Phnom Penh, Cambodia
{meily.oeng.2018,chap.chanpiseth,srun.sovila,khim.chamroeun}@rupp.edu.kh

***Abstract* — The development of efficient credit scoring algorithms is essential given the exponential rise of consumer credit and the large volume of financial data. In recent years, researchers have developed sophisticated credit scoring models using statistical, machine learning, and artificial intelligence (AI) techniques to provide banks and other financial institutions with a tool to make better decisions in credit risk assessment. The goal of this study is to propose a credit scoring model using deep learning, i.e., LSTM, to predict customer behavior on defaulted loan payments. The data used in this research was obtained from a public machine learning repository containing both customer demographic information and sequential data of their loan payment history. Following our observation, the dataset contained a class imbalance due to a minor class of default loan payment. Therefore, the Synthetic Minority Oversampling Technique (SMOTE) was introduced to oversample the minor class. Based on our experiment, the proposed credit scoring model achieved approximately 85 percent in terms of accuracy.**

***Keywords*— Credit Scoring, Default on Loan Payment, Deep Learning, LSTM, Sequential Data**

I. INTRODUCTION

Numerous empirical evidences demonstrate that the most significant loss suffered by any financial organization is primarily caused by an inefficient and inadequate credit risk management procedure, leading to bankruptcy and insolvency within the financial institution[1].

An effective credit scoring model is widely understood as a valuable tool for improving financial inclusion by providing responsible and sustainable credit access for individuals and businesses. Being able to access financial inclusion provides immense potential to contribute to the economic growth of the world[2].

In recent years, there has increasingly caught the interest from numerous research studies to propose automatic credit scoring models. The researchers have introduced sophisticated techniques, from traditional statistical modeling to modern machine learning and artificial intelligence approaches. The machine learning algorithms include random forests, gradient boosting, deep neural networks, etc.

The theoretical background of credit scoring will be discussed, from its definition, the importance of credit scoring models used in credit risk assessment, and the evaluation procedures incorporated to validate the credit scoring models. Finally, the proposed credit scoring methodologies, including statistical methods and machine learning techniques to learn the credit scoring models, will be described in detail.

- *Credit Scoring Model*

Credit scoring is a method used to predict the probability that a loan applicant, existing borrower, or counterparty will default or become delinquent. It is widely used for consumer lending, credit cards, and mortgage lending.

It has become an essential tool in the credit management process for banks and financial institutions since they are well aware of the numerous risks they have been facing, particularly those related to the granting of loans to customers. Banks and other financial institutions collect customers' information, thoroughly analyze their data, and then make a final decision on whether to accept or reject that loan. In short, the goal of the credit scoring model was introduced to assist financial organizations in deciding whether or not to provide lending to a customer.

The credit scoring models provide an output in the form of probability. Therefore, the decision maker compares the estimated output of the model against a predefined threshold value to validate a loan application. The higher the outcome of the prediction on the loan applicant compared to the threshold value, the customer is considered a credit-worthy customer who is likely to repay financial obligations. Otherwise, the loan application is considered more at risk and less likely to repay the loan. The adoption of credit scoring models can have a huge impact on profitability for banks and financial organizations, where a small fraction of a percentage improvement in credit risk assessment will have a tremendous impact on their financial health.

II. RELATED WORKS

To date, many credit scoring algorithms have been proven to successfully evaluate and improve credit risk management in research studies. There have been many credit scoring methodologies that have been proposed in recent years to address the credit score challenges. The most extensively used credit scoring models are logistic regression and decision tree. Those credit scoring methodologies could be classified into two categories, such as application scoring models and behavioral scoring models.

Application scoring is done at the time of application and estimates the probability of default using application data such as amount of loan, type of loan, maturity of loan, guarantees,

and collateral value. whereas behavioral scoring is used after the loan has been granted and is generally used to monitor an individual's probability of default over some period of time based on historical payment performance or the customer's behavior on transaction records and/or online operations.

The authors proposed an application scoring model for the retail banking sector that used several statistical techniques such as Logistic Regression, Discriminant Analysis, and Probit Analysis with application data from Askari Bank Limited in Pakistan[3]. The accuracy of the proposed methods varies from 81 to 85 percent, respectively. This indicates that the performance of the statistical methods is still limited and needs to be improved.

In 2018, a group of researchers from China proposed a novel credit scoring model for peer-to-peer lending using the attention mechanism with LSTM[2]. The dataset used in this study was collected from the online operations of multiple lenders, and this dataset is known as behavioral data. The proposed model in this research outperformed other traditional models discussed in [1]. However, there is a drawback to this model since the authors took into consideration only the sequential online operation events. Ignoring time intervals in the dataset may negatively impact the performance of a model. As can be observed in the dataset, the gaps between sequential events vary from one lender to another.

III. METHODOLOGY

A. Dataset Collection

Every machine learning algorithm relies heavily on high-quality data. There are various ways to collect data. However, this model needs financial data, which includes sensitive consumer data that cannot be made public.

Credit Bureau Cambodia [4] is the top supplier of financial data, analytical tools, and credit reporting services to businesses, consumers, and financial institutions. A personal credit report of CBC is a summary of a borrower's credit payment history that is gathered from banks, microfinance organizations, and other significant financial institutions. A strong credit payback history will make it simpler for consumers to qualify for loans and receive credit because all lenders will review the consumer's personal credit file to determine their creditworthiness before making a choice.

The K-score[5] model of CBC is a number that summarizes all the information from the credit report into a single score that has four categories that include all types of data, such as financial data, demographic data, and customer behavioral characteristics. On the other hand, this research is mainly focusing on behavioral scoring models that only require behavioral data like the previous payment schedule.

TABLE I
DATASET FEATURE DESCRIPTION

Dataset Feature Description	
Limit Balance	(NT Dollar): 10,000 - 1000,000
Age	(Years old): 21-79
Amount of bill statement	(NT Dollar)
Amount of previous payment	(NT Dollar)
Marital Status	1 = married; 2 = single; 3 = others
Past Payment Status	[-1] indicate pay duly, whereas [1, 2, 3, 4, ..., 9] payment delay from1 to 9 months, respectively
Gender	1 = male; 2 = female
Education level	1 = graduate school; 2 = university; 3 = high school; 4 = others

The selected dataset was the only one that matched the CBC personal report feature that described the consumer's previous loan payment history and was publicly available. The dataset used in this study is a public Taiwanese non-transactional credit card dataset that includes consumer default payments and has been published on the University of California, Irvine's official website called "The UCI Machine Learning Repository"[6]. This dataset contains 30,000 records, which is sufficient to assess the performance of the proposed model. There are 23,364 non-default payments, while there are 6636 default payments (the proportion of default payments in the dataset is 22 percent).

The variables are divided into two groups: numerical and categorical. The examples of the first are: the number of given credits; the aged payment statement; and the past payment bill. The second group includes variables such as gender, education, and marital status. This research employed a binary variable, default payment (Default = 1, Non-default = 0), as the response variable.

B. Data Preprocessing

The negative number's presence in the dataset, which surely is an outlier. The variable that contains the most outliers is the previous payment that has been made in the last 6 months. The mean of these is approximately 5,000, but some values are far above the mean, and some even reach more than one million.

To use machine learning to solve our prediction, we must first manipulate the data to fit the algorithms' input format. To do so, we cleaned each of the datasets separately before combining them. Besides the missing value, there are many other problems in this dataset, such as noise and outliers. Noise occurs in two categorical variables, namely marital status and education level, and might be caused by typing errors during the data collection phase. After our analysis, we chose to deal with this problem by dropping the record that contains this error, both noise and outliers.

Some classifiers in deep learning require input values that range from 0 to 1 and in vectors of real numbers. However, the datasets contain inputs that hold values that vary in range. In order to avoid bias and accordingly feed the algorithm with data within the same interval, data should be transformed from a different scale of value. In order to achieve this, attributes should be normalized to values in the range of 0 and 1, using the normalization rule and standardization rule[7].
Besides that, as it was mentioned above, the dataset that was chosen to train this model has two classes, which are non-default payment and default payment. The proportion of non-default payments is 77.88 percent, while the default payment is only 22.12 percent, which is leading to a class imbalance problem. The Synthetic Oversampling Technique (SMOTE) was used to solve this problem by randomly increasing minority class examples by replicating them[8].
Finally, the sequential data was reshaped into a three-dimensional array of shapes (number of customers, number of months, and number of features) before being fed to our model.

C. Long Short-Term Memory (LSTM)

Long-Short-Term Memory Networks, or "LSTMs," are a type of RNN[9] that can learn long-term dependencies [10]. LSTMs are specifically developed to prevent the problem of long-term dependency. They don't have to work hard to remember knowledge for lengthy periods of time; it's like second nature to them. All recurrent neural networks are made up of a series of repeated neural network modules. This repeating module in ordinary RNNs will have a relatively simple structure, such as a single Tanh layer. LSTMs also have this chain-like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way (Fig. 1).
A distinctive structural unit and three kinds of unique gate structures are designed in the LSTM network, one of which is an input gate that determines how many cell states must be stored. An output gate that determines how many cell states must be delivered to the next cell, and a forget gate that determines how much data must be erased. Internal states exist at two of these gates.

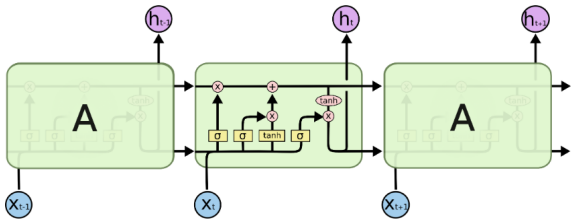


Fig. 1 LSTM architecture[11]

The shape of the activation function is crucial and can have a major impact on the neural network's efficiency. The information that passes through the unit is added or removed selectively. The sigmoid function is used to implement the gate structure. The sigmoid value, which goes from 0 to 1, indicates how much data can be passed through. In feedforward neural networks, the sigmoid function is a non-linear activation function. It is a bounded monotonically increasing real function, defined for all real input values as given by the following sigmoid function equation:

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (1)$$

The hyperbolic tangent function is the default activation function of the LSTM. The smooth antisymmetric hyperbolic tangent function, tanh, has a range of values of [-1, 1]. The tanh function's equation is as follows:

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \quad (2)$$

The Tanh function's key benefit is that it generates zero-centered output, which aids the back-propagation process. The following is a detailed description of how an LSTM cell works: To begin, the LSTM unit uses the forget gate to process the information from the previous memory state in order to identify which information should be erased from the memory state.

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \quad (3)$$

The LSTM unit then decides what data should be kept in memory. The input gate, on the other hand, selects which data to update. A tanh layer, on the other hand, updates the candidate vector.

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b)_i \quad (4)$$
$$C_t = tanh(W_c.[h_{t-1}, x_t] + b_c) \quad (5)$$

In the next step, the LSTM unit combines the two pieces previously mentioned to update the memory state.

$$C_t = f_t * C_{t-1} + i_t * C_t \quad (6)$$

Finally, the output gate regulates the memory state that must be output.

$$O_t = \sigma(W_O.[h_{t-1}, x_t] + b_O) \quad (7)$$
$$h_t = O_t * tanh(C_t) \quad (8)$$

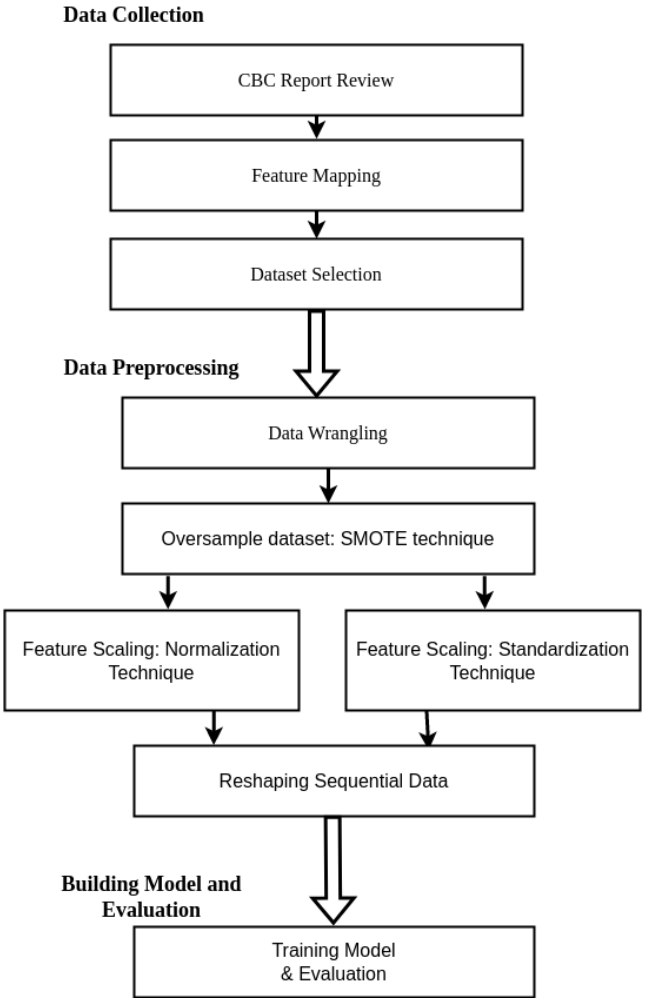


Fig. 2 Methodology Overview

D. Proposed Method

The proposed model's framework is shown below(Fig. 2). We can thoroughly investigate the model's performance using the proposed approach and draw trustworthy results. Firstly, our model has two inputs: sequential data and application data. The latter is a 3 dimensional array and is our main input that feeds into the LSTM model. It is then flattened into a 2 dimensional shape in order to concatenate with the additional input, which is application data. Lastly, dense layers are applied as the last layer of our model. To validate the model, we split our dataset into two: the training set and the test set, 80% and 20%, respectively. The model is evaluated using the 20% of the records in our selected dataset and we achieve approximately 85 percent in terms of accuracy.

TABLE II
MODEL HYPERPARAMETER

Hyperparameter	
Activation Function of LSTM	Tanh
Activation of output layer	Sigmoid
Loss Function	Binary cross entropy
Optimizer	Adam
Dropout	0.1

IV. RESULT AND DISCUSSION

During the preprocessing phase, we chose to experiment with two types of feature scaling: normalize and standardize. For one main reason, we acknowledge that the most effective technique to scale our dataset is the standardization rule. After doing exploratory data analysis, we acknowledge that our dataset contains a lot of outliers in all numerical variables. The main reason we were unable to apply the normalization rule is that it is influenced by outliers, but standardization is not. It shows that the percentage of normalization is inaccurate compared to the number of standardization rules. As a matter of fact, we decided to use standardization techniques for this project.

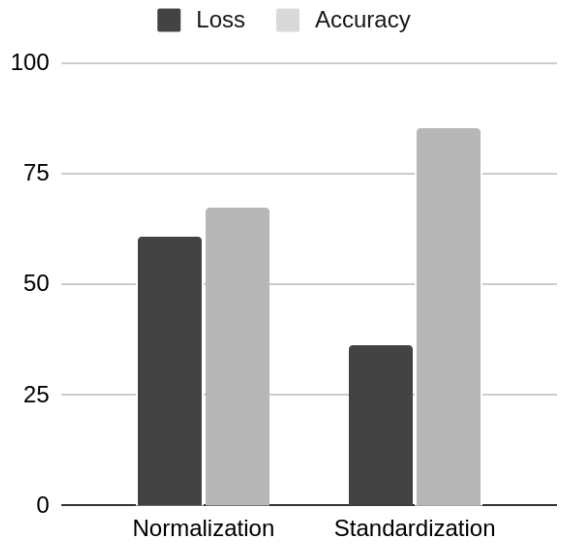


Fig. 4 Loss and Accuracy comparison

After the experimentation, we realized that dropping all of the outliers could also affect our model performance. This is due to our dataset containing too many outliers, so when we drop all of them, the dataset remains in a very small amount of the record, which is not enough to train our model.

The performance is according to the evaluation in the test set that we split from one-fifth of our whole dataset. Our model reaches its highest performance, which is approximately 85% in terms of accuracy.

TABLE III
MODEL PERFORMANCE

	Accuracy	Recall	Precision	F1 Score

LSTM Model	85.11%	78.81%	89.31%	83.73%

V. CONCLUSIONS

The importance of credit scoring for measuring and reducing bank losses is emphasized in this paper. After doing research, it was discovered that the LSTM model has the highest accuracy in predicting late fees and missed repayments, making it the ideal option for banks. In this research, the LSTM model was proposed and verified on a non-transactional open dataset. Banks can use the model's prediction not only as a binary output (whether a customer would miss a payment in the following month), but also as a score for each client. The LSTM predicts if a client will become insolvent in the coming month. Management is responsible for establishing thresholds at which the bank classifies this user as high or medium risk, with accompanying consequences for the user. To put it another way, the LSTM model's outputs can be used to categorize clients into risk groups.

VI. FUTURE WORK

This research has a few areas that can be further developed. One obvious area is to improve the performance by trying this model with other datasets, especially from Cambodian financial organizations, to be effective for credit rating based on the available Cambodian loan applications and financial features. Another future project that we want to carry out is to find out what techniques CBC uses to calculate their K-score model, and we also want to compare my model's performance with K-score.

ACKNOWLEDGMENT

This study would not have been possible without my supervisor, Mr. Chap Chanpiseth. I would like to offer my heartfelt gratitude to him for his genuineness and encouragement, which I will never forget. I am grateful for the tremendous possibilities he provided for me to progress professionally and for the extraordinary experiences he planned for me. Additionally, I would like to express my sincere gratitude to Dr. Srun Sovila and Dr. Khim Chamroeun for their guidance and assistance with the thesis research and academic journey.

REFERENCES

[1] Klepkova Vodova, Pavla. (2003). Credit Risk as a Cause of Banking Crises.

[2] Vidal, Maria Fernandez, and Fernando Barbon. 2019. "Credit Scoring in Financial Inclusion." Technical Guide. Washington, D.C.: CGAP.

[3] Hussain, Arif & Khan, Muhammad Imran & Rehman, Shams & Khattak, Aaiya. (2019). Credit Scoring Model for Retail Banking Sector in Pakistan. 14. 153-161.

[4] Wang, Chongren & Han, Dongmei & liu, Qigang & Luo, Suyuan. (2018). A Deep Learning Approach for Credit Scoring of Peer-to-Peer Lending Using Attention Mechanism LSTM. IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2887138.

[5] *K-Score*. Available at: https://www.creditbureau.com.kh

[6] Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.

[7] Ahsan, M. *et al.* (2021) "Effect of data scaling methods on machine learning algorithms and model performance," *Technologies*, 9(3), p. 52. Available at: https://doi.org/10.3390/technologies9030052.

[8] Xu, J., Lu, Z. and Xie, Y. (2021) "Loan default prediction of Chinese P2P market: A machine learning methodology," Scientific Reports, 11(1). Available at: https://doi.org/10.1038/s41598-021-98361-6.

[9] Beysolow II, T. (2017) "Recurrent neural networks (rnns)," Introduction to Deep Learning Using R, pp. 113–124. Available at: https://doi.org/10.1007/978-1-4842-2734-3_6.

[10] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

[11] Olah, C. (2015) *Understanding LSTM networks*, *Understanding LSTM Networks -- colah's blog*. Available at: https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Attendance Marking System using Face Recognition Technique based on SVM and PCA

Vysal Vothy, Chamroeun Khim, Sovila Srun, Chanpiseth Chap

Department of Information Technology of Engineering, Royal University of Phnom Penh
Phnom Penh, Cambodia

{vothy.vysal.2018,khim.chamroeun,srun.sovila,chap.chanpiseth}@rupp.edu.kh

Abstract— **Face recognition is one of the significant techniques in computer vision that works by outputting the information about the face that was detected in the image. In this research, we proposed developing a face recognition system to automatically detect students' faces and record their attendance. The purpose of this study is to collect the students' attendance by implementing a face recognition application. The efficient face recognition method known as Principal Component Analysis (PCA) has been used in this work. In addition to this, we proposed machine learning algorithms such as Support Vector Machine (SVM) and Naive Bayes to classify the detected faces to match their labels. The output of our model can be exported as an excel file for further use. After our experiment with twenty two faces we observed that Support Vector Machine (SVM) can produce a better performance, 88 percent in terms of accuracy.**

Keywords— **Face Recognition, Face Attendance, Principal Component Analysis, Support Vector Machine, Naïve Bayes**

I. INTRODUCTION

Rapid growth of technology in recent years contributes to the increase of electronic devices manufacturing. Those electronic devices such as smartphones, laptops, desktops and other devices are playing an important role in business, communication, healthcare, scientific, education etc. With the help of technology, human life become more comfortable.

In addition to the increase in the demands of the electronic devices, numerous software and systems have been developed to fulfill the need of daily usage and business for instance Microsoft Office, YouTube, Facebook, etc. Those software and systems do not address all the needs, then Computer Vision are coming into technology's world as embedded software. Face Recognition is one of the technique in the Computer Vision that was introduced to solve real world problem for instance screen unlock in smartphone using Face Recognition. Furthermore, many researches have implemented face recognition in various domains. In education domain, face recognition technique were introduced to track students' attendance records. Traditional system for attendance tracking is remarkably time-consuming teachers need to manually collect the students' attendance. For this reason, many researchers were motivated to address these traditional ways. Previous studies [1], [2] discussed the challenges in automatic face recognition. Those include illumination in image, head posture, variation and postures found in faces, stages of life of a human being and multiple face in one frame which cause the low accuracy in face recognition.

In this research, we propose to develop the automatic system to record students' attendance in real time using face recognition. We adapt the Principal Component Analysis (PCA) for feature selection and representation. Most significantly, we use the support vector machine and Naïve Bayesian for face recognition and compare the recognition performance of both two techniques.

II. LITERATURE REVIEW

Until now, many researches have been conducted in face recognition. Several methods commonly used are Local Binary Pattern Histogram (LBPH), Fisher Face, and Eigen Face.

In [1] has proposed the Face-Recognition by applying LBPH in their study and got 77 percent of accuracy. Author [2] has been used Eigen Face and got 86 percent of accuracy. Besides the different methods that were used in both of these studies, we can see some similar strategies that they used in their experiment. They had resized the cropped face image into a size of 110 x 110. Additionally, we can notice the difference in usage between LBPH and Eigen Face. With the Eigen Face technique, the size of the face image between training and recognition needs to be the same, while Local Binary Pattern Histogram (LBPH) techniques can have a different size of the image between training and recognition. Furthermore, there are several researches have been conducted to analysist the accuracy between three of these method mentioned like [3] and [4] have analyst these three method and get the result as shown in TABLE I.

TABLE I
The three methods used in face recognition

Research No.	Face Recognition Algorithm	Accuracy		
		PCA	LBPH	Fisher Face
[3]	Face Recognition An Engineering Approach	95%	91%	95%

[4]	Comparative Analysis of Face Recognition Methodologies and Techniques	90%	75%	88%

According to the four references that we mentioned, we can see [1], [2] did the automated face recognition attendance in real time which is similar to our study, but both of them used the library of Open Computer Vision (Open CV) for training and recognition. So, it's a challenge for us to customize and apply machine learning to the dataset. On the other hand, even [3], [4] get the high accuracy with Principal Component Analysis (PCA), but their research did not apply with real time video.

III. METHODOLOGY

3.1. Dataset Overview

We gathered twenty two people for our study. We took fifty pictures of each person from various angles and saved them in a folder labeled with their names. For instance, first person will be taken fifty images and then stored in folder's name first person.

As a result, we got one thousand one hundred face images from those twenty two people as the dataset.

3.2. Architecture

In this section, we will show the architecture that we implemented to develop this application.

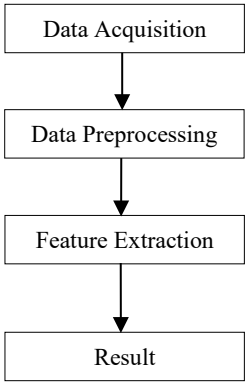


Fig.1. Architecture

3.3. Experimental Design

- Data Acquisition: It is the process by which we photograph people and save them in a folder labeled with their names. Furthermore, each of those images was saved in JPG format. At this point, each person will be photographed fifty times.

Fig.2. Data Acquisition

- Data Pre-processing: We crop the face from the image during this process. To crop the face, we must first determine the location of the face using a face detection algorithm. Haar-Cascade was the face detection algorithm used in our study. In this section, all images must be converted to greyscale before applying Haar-Cascade to obtain the face image. The cropped face image was then resized to 100 x 100.

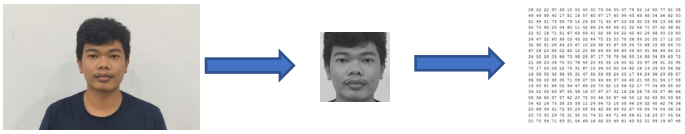


Fig. 3. Data Pre-processing

- Feature Extraction: We take the array of the cropped face and apply Principal Component Analysis (PCA) to that face vector to obtain the Eigen Face.

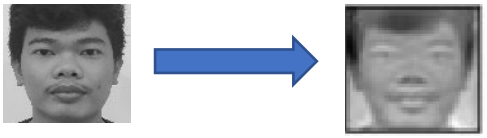


Fig. 4. Feature extraction

- Result: In both training and recognition, Support Vector Machine (SVM) and Naïve Bayes will be used. If the system detects human faces in the images, it will record that information and save it in an excel file.

  We used a webcam and live lamp to improve the image quality and brightness of the environment.
- Webcam: FANTECH LUMINOUS C30 is the name of the webcam that we used for this project. This webcam's specification are as follows:

Fig. 5 FANTECH LUMINOUS C30 [5]

  - Video resolution: 2560 x 1440p
  - Field of view: 106°
  - Frame rate: 2k/ 25 fps (HD/25fps for Ubuntu)
  - Microphone build-in microphone
  - Plug type: USB 2.0
  - Cord Length: 1.4m
  - Dimension: 73 * 26 * 33 mm (camera)
  - Weight: 87 gr
- Live Lamp: This component was used when photographing a user for the dataset. It is used to adjust the illumination in the environment in order to produce a high-quality image. The live lamp specification is detailed below:

Fig. 6 F-450 18" Ring Light With Stand [6]

- Source type light: LED
- The main field of application: direct lamp fill beauty light
- Light source power: 24w voltage: 2V 5A
- Tripod: 110 cm
- Dimmable: Yes
- Led Weight Light: 400g Tripod
- Weight: 900g
- Led light Size: 25.5 CM
- Use: photography / Live Colour temperature: 2700k-5500k

3.4. Principal Component Analysis

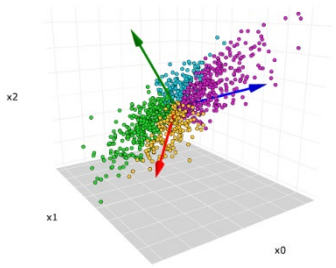
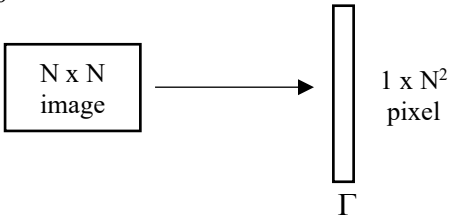


Fig. 7 Principal Component Analysis [7]

Principal Component Analysis is the mathematic statically was used to reduce the large data set to small dataset without losing the information. Principal Component Analysis (PCA) has been implemented through various step which described below:

- Assume M is the total of face image in the dataset. Since all of the images has the same size of N x N. Each of face dimension has been converted into face vector with size of 1 x N2 which denoted as Γ. In our case face image has the size of 100 x 100 then converted to 1 x 10000



- Then every value of the face vector has been standardized in the range [0-1] by divide each value of the vector with 255.
- Calculating the average face.
  The average of face has been calculated by:

$$\Psi = \frac{1}{M}\sum_{n=1}^{M} \Gamma_n \qquad (1)$$

  Each face differs from the average by the vector:

$$\Phi i = \Gamma_i - \Psi, i = 1, \dots, M. \qquad (2)$$

- Calculating the covariance matrix.
  Covariance is defined by:

$$C = \frac{1}{M}\sum_{n=1}^{M} \phi_n \phi^T = AA^T, \qquad (3)$$

$$where\ A = [\phi_1, \phi_2, \dots, \phi_M]\ is\ N^2\ x\ M\ matrix$$

- Calculating the eigenvectors and eigenvalues of the covariance matrix.
  Finding eigenvectors (v_l) from the covariance matrix is a time-consuming computational task. We can construct the M by M matrix

$$L = A^T A \qquad (4)$$

  because M by M is far less than N^2 by N^2 where

$$L = L_{mn} = \phi_m^T \phi_n = A^T A$$

  .
  These vectors (v_l) determine linear combinations of the M training set face images to form the eigenfaces u_l.

$$u_l = \sum_{k=1}^{M} V_{lk}\, \phi_k, l = 1,2, \dots, M \qquad (5)$$

- Selecting the principal components.
  Each of the original images should be projected into eigenspace. This produces a vector of weights representing each eigenface's contribution to the reconstruction of the given image.
  $\omega_k = u_k^T(\Gamma - \Psi) \quad \Omega^T = [\omega_1, \omega_2, \dots \dots \dots, \omega_M]$, (6)
  where u_k is the k^{th} eigenvector and ω_k is the k^{th} weight in the vector $\Omega^T = [\omega_1, \omega_2, \dots \dots \dots, \omega_M]$

The result of Eigen Face shown in Fig.8.

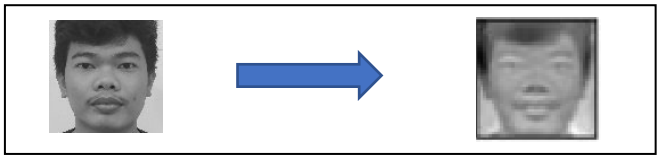


Fig. 8 Eigen Face

3.5. Support Vector Machine (SVM)

SVM is the abbreviation of Support Vector Machines are a set of supervised Learning methods used for classification, regression and outliers' detection. It will create the linear line in the graph. The line will represent the class name. Every input image will be calculated as the point and will plot somewhere

in the graph. And then it will calculate the origin between point to the line by consider the nearest one to the line of class name will be member of that class and then output the result.

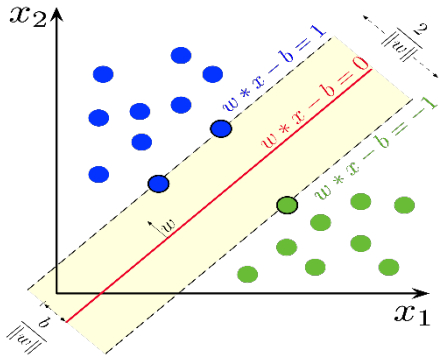


Fig. 9 A Separable Classification [8]

The main goal of SVM is to maximize the distance in order to give some space to the hyperplane equation. The hyperplanes equation denoted as:

$$(w.x) + b = 0\, w \in R^N, b \in R, \tag{7}$$

Corresponding to decision functions:

$$f(x) = sign((w.x) + b, \tag{8}$$

where $w = \sum_i v_i x_i$ *in term of a subset of training patterns that lie on the margin* (*see in Fig.* 9.)

In Fig.10. Show the basic idea of SVM, which is to map the data into some other dot product space (called the feature space) F via a nonlinear map.

$$\Phi: R^N \rightarrow \mathrm{F}, \tag{9}$$

This decision surface has the equation:

$$f(x) = \sum_{i=1}^{l} v_i \,.k(x_i, x) + b \tag{10}$$

The kernel function used in the experiments are linear defined as Linear Kernel:

$$k(x_i, x) = x_i.x \tag{11}$$

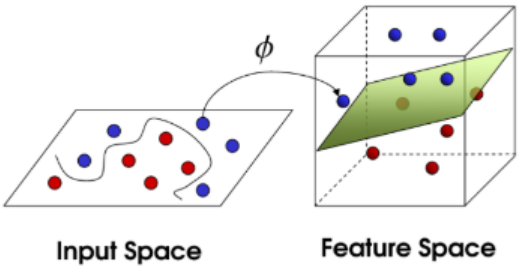


Fig. 10 Matching Point to Feature Space [9]

3.5. Naïve Bayes

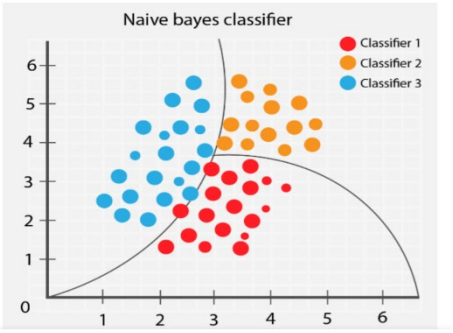


Fig. 11 Naïve Bayes Classifier [10]

Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification tasks. Bayes' Theorem is a simple mathematical formula used for calculating conditional probabilities. Conditional Probability is a measure of the probability of an event occurring given that another event has (by assumption, presumption, assertion, or evidence) occurred. Bayes theorem equation:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{12}$$

Where P(A|B) is the probability of observing event B given that A is true. P(A) and P(B) are probabilities of observing A and B. In Bayes rule we go from P (X | y) that can be found from the training dataset to find P (y | X). To get this, we need to replace X and y with A and B. Where X is the known variable and y is the unknown variable. Since the X is given as:

$$X = (x_1, x_2, \ldots., x_n)\ where\ x_1, x_2, \ldots, x_n \tag{13}$$

represen the features [*In our case is* 100 *features*]

By substituting for X and expanding using the chain rule we get,

$$P(y \mid x_1, x_2, \ldots., x_n) = \frac{1}{z}\, p(y)\, \mathbb{\Pi}_{i=1}^{n} p(x_i \mid y), \tag{14}$$

where the evidence $Z = p\,(x) = \sum p(y)p(x\,|y)$ *is* a scaling factor dependent only on $x_1, x_2, \ldots., x_n$, *that is a constant if the values of the feature variables are known* .

Meanwhile, independent probability is the influence of all data features on each class y. For continuous data, Nave Bayes can be described as follows.

$$P(x = v|y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma^2}} \tag{15}$$

3.7. Haar Cascade

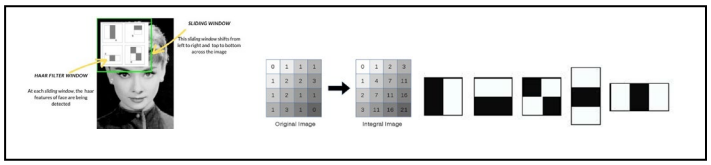


Fig. 12 Haar Cascade [11]

Haar-Cascade is an Object Detection Algorithm used to identify faces in an image or a real time video. The algorithm uses edge or line detection features proposed by viola and Jones in their research paper "Rapid Object Detection using a Boosted Cascade of Simple Features" published in 2001. The algorithm is given a lot of positive images consisting of faces, and a lot of negative images not consisting of any face to train on them.

IV. Conclusions and Future Work

After doing this research, we can see the different performance between Support Vector Machine (SVM) and Naïve Bayes. With Support Vector Machine (SVM) generated lower accuracy than Naïve Bayes but the model that Support Vector Machine (SVM) output is more balance than Naïve Bayes. Based on that result, we decide to use Support Vector Machine (SVM) for training model in our research. As the results we can develop our system which allow users to perform some task like record the student's attendance which use less time in checking students' attendance. But this system can take only one students' per time. So, we will improve our research more to have ability to take multiple students' in one time. In addition, we will find out the technique to make the balance of illumination in image to get the high accuracy.

Acknowledgment

I would like to express my sincere gratitude to my professor, Dr. Khim Chamroeurn, for giving me such a golden opportunity to work on this wonderful project. Additionally, I would like to express my sincere gratitude to Dr. Srul Sovila and Mr. Chap Chanpiseth for his guidance and assistance for the research and academic journey.

References

[1] Sitriprajna Panda, Swati Sucharita Barik, Sasmita Kumari Nayak, Aeisuriya Tripathy, Gourav Mohapatra: Human Face Recognition using LBPH (International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volumne-8 Issue-6, March 2020).

[2] Khem Puthea, Rudy Hartanto, Risanuri Hidayat: The Attendacce Marking System based on Eigenface Recognition using OpenCV and Python (Journal of Physics: Conference Series 1551(2020) 012012 doi: 10.1088/1742-6596/1551/1/012012).

[3] Ghahramani, Farshad, "Face Recognition: An Engineering Approach" (2015). Master's Theses. 4635. DOI: https://doi.org/10.31979/etd.zb9p-5z5h

[4] Farwa Abdul Hannan, Zainab Khalid, Ammar Rafiq: Comparative Analysis of Face Recognition Methodologies and Techniques (NFC-IEFR JOURNAL OF ENGINEERING & SCIENTIFIC RESEARCH ISSN: 2222-1247, VOL. 04: DECEMBER, 2016).

[5] Luminous C30 2K QHD webcam: Fantech (2022) Fantech World. Available at: https://fantechworld.com/luminous-c30/ (Accessed: December 20, 2022).

[6] Billah, M. (2022) F-450 18" Ring light with stand, Billzumla. Available at: https://billzumla.com/shop-2/studio-equipment/ring-light/f-450-18-ring-light-with-stand/ (Accessed: December 20, 2022).

[7] Cheng, C. (2022, March 22). Principal Component Analysis (PCA) explained visually with Zero math. Medium. Retrieved December 20, 2022, from https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d

[8] Support Vector Machine (2022) Wikipedia. Wikimedia Foundation. Available at: https://en.wikipedia.org/wiki/Support_vector_machine (Accessed: December 20, 2022).

[9] Radhika (2020) Mathematics behind SVM: Math behind support vector machine, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2020/10/the-mathematics-behind-svm/ (Accessed: December 20, 2022).

[10] Chaudhuri, K.D. (2022) Building naive Bayes classifier from scratch to perform sentiment analysis, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2022/03/building-naive-bayes-classifier-from-scratch-to-perform-sentiment-analysis/ (Accessed: December 20, 2022).

[11] Jaiswal, A. (2022) Object detection using Haar Cascade: Opencv, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2022/04/object-detection-using-haar-cascade-opencv/ (Accessed: December 20, 2022).

High Availability of Proxy Server on PfSense

Choeun Kongkea, NhemThayheng, Hul Sovannaroth
Cambodia Academy of Digital Technology
Phnom Penh, Cambodia
kongkea.choeun@student.cadt.edu.kh
thayheng.nhem@cadt.edu.kh
Smart Axiata, Phnom Penh, Cambodia
hul.sovannaroth@smart.com.kh

***Abstract*— Many of our daily activities in this decade of the digital world are reliant on the internet, and various forms of communication, entertainment, financial and work-related tasks are accomplished online. Although most of the internet is private and secure, there are occasionally vulnerable websites or web pages too. In order to control or filter internet access in organizations or workplaces, we have to use a proxy server. Moreover, there are many open-source proxy servers that provide security, functionality, and privacy depending on our needs and company policy. Furthermore, to ensure the proxy server elimination a single point of failure and the service is still available. Hence to respond to these problems, by implementing high availability of pfsense for the proxy server which is the open software firewall and router. In order to achieve high availability for pfsense, we have to combine some features such as CARP, Pfsync, and XMLRPC.**

***Keywords*— Pfsence, Proxy server, Common Address Redundancy (CARP), Pfsync, High availability.**

I. INTRODUCTION

In this decade, digital landscape, many of our daily activities rely on the internet, and various forms of communication, entertainment, financial and work-related tasks are accomplished online. The internet is mostly private and secure but it can be the insecure website or web pages. Therefore, cyber attackers can attack an organization's business or threaten people's privacy. That can disrupt e-commerce, cause the loss of business data. To prevent all these cases and also improve security, secure employee's internet activity from hackers trying to snoop on them, and control the websites employees and staff access to the internet by using proxy server that offers security, functionality, and privacy for our needs. A router or system that acts as a gateway for users to access the internet is known as a proxy server.[1] Additionally, there are many open-source firewall that provide a solution designed to act as a guard between external and internal networks. There are two top open-source firewalls that can act as proxy server and support many features for companies or organizations to consider in the new digital world.

The OPNSense firewall is easy-to-use, free, and ideal for infinite scalability. This open-source project promises best-in-class virtual private networking, intrusion detection, and a powerful firewall with support for both IPv6 and IPv4 live view on passed and blocked traffic. Multi-WAN capability is included with hardware failover, state synchronization, and intrusion detection.[2]

However, PfSense is a free open-source customized distribution of FreeBSD tailored for use as a firewall and router entirely managed by an easy-to-use web interface. This web interface is known as the web-based GUI configurator or WebGUI for short. Mainly is used as a router and stateful firewall. It has many packages extend its capabilities such as Squid3 package as a proxy server that cache data and SquidGuard, redirector and access controller plugin for squid3 proxy server. [3]

Therefore, even in the organizations or workspace use a proxy server is not enough. We need to elimination a single point of failure. If one proxy server in our infrastructure down, the another will take over. In this paper, to respond the issues, PfSense provide the high availability that have combine some features such as CARP,Pfsync,and XMLRPC.

II. METHODOLOGY

Main Concept

High availability is a feature that provides redundancy and fault tolerance. It is a number of connected devices processing and providing a service. Its goal is to ensure this service is always available even in the event of a failure.

In firewalls and other similar devices, the high availability feature is a mechanism to keep the state of devices synchronized with each other as well as being able to detect a failure so that if a failure did occur active devices would know about this and be able to take on the processing load from the failed device. By using pfSense software which is the one of very few open-source solutions offering enterprise-class high availability capabilities with stateful failover, allowing the elimination of the firewall as a single point of failure. There are several features to achieve high availability such as CARP, XMLPRC, pfsync.

CARP was created by OpenBSD developers as a free, open redundancy solution for sharing IP addresses among a group of network devices. CARP is a type Virtual IP address (VIP) shared between nodes of a cluster. One node is the master and receives traffic for the IP address, and the other nodes maintain backup status and monitor for heartbeats to see if they need to assume the master role if the previous

master fails. Since only one member of the cluster at a time is using the IP address, there is no IP address conflict for CARP VIPs.[4]

PfSense XML-RPC config Sync

To make the job of maintaining practically identical pfSense software nodes easier, configuration synchronization is possible using XML-RPC. When XML-RPC Synchronization is enabled, settings from supported areas are copied to the secondary and activated after each configuration change. XMLRPC Synchronization is optional, but maintaining a cluster is a lot more work without it.[5]

PfSync is the feature to synchronize the firewall state table between nodes of a cluster. When the state table change on the Master node, it will send to the Backup node over the Sync interfaces, and vice versa. By default, State Synchronization with pfsync uses multicast. Furthermore, high availability can still work without State Synchronization, but it will not be smooth. If no State Synchronization, while a node fails and another node takes over but the user connection would be dropped. [6]

Flow Topology

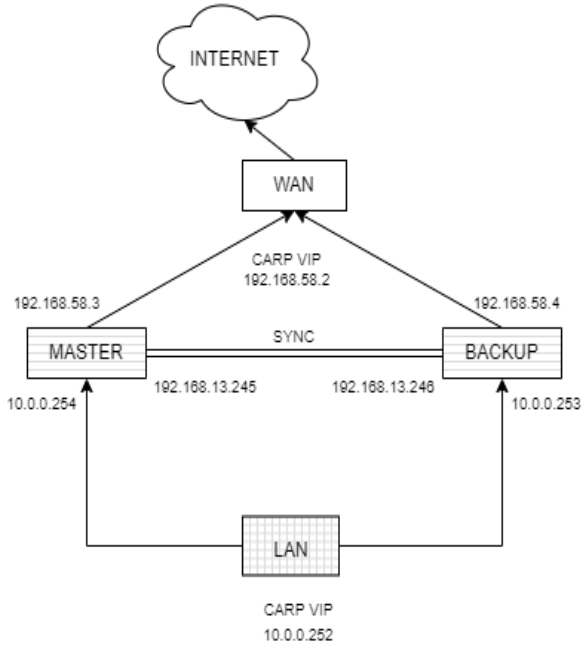


Figure.1 Topology

In order to demonstrate the process of the of research, topology plays an important role to facilitate people understanding the research. Figures 1 show that there are two of pfsense act as proxy server, one is the master, another is the backup. The master and backup each have identical connections to the WAN and LAN, and a crossover cable between them to connect the Sync interfaces. For the Sync interface is used for synchronization with pfsence feature called XML-RPC config sync and also exchange the state table between the master node and the backup nodes. Moreover CARP VIP (Virtual IP address) is use in LAN and WAN interfaces. High availability cluster of pfsense using CARP needs three IP address in each subnet and it separates unused subnet for the sync interface. For WANs and LAN, are required a /29 subnet each. That's mean one IP address is used by each node the master node and the backup node, add a share CARP VIP address for failover. However, the synchronize interface only requires one IP address per node.

III. IMPLEMENTATION

In order to implement the high availability of proxy server on pfsense. The first task is determined IP address assignments.

IP address	Usage
192.168.58.3/24	Pfsense Primary Wan IP address
192.168.58.4/24	Pfsense Secondary Wan IP address
192.168.58.2/24	CARP shared IP address

Figure.2 Wan IP Addresses Assignments

IP address	Usage
10.0.0.254/24	Pfsense Primary Lan IP address
10.0.0.25324	Pfsense Secondary Lan IP address
10.0.0.252/24	CARP shared IP address

Figure.3 Lan IP Addresses Assignments

After determined IP address, set up the Sync interface on the cluster nodes must be configured. Use Sync IP addresses Assignment lists below on each Sync interface with the appropriate IPv4 address value.

IP address	Usage
192.168.13.245/24	Pfsense Primary Sync IP address
192.168.13.246/24	Pfsense Secondary Sync IP address

Figure.4 Sync IP Addresses Assignments

State synchronization using pfsync must be configured on both the primary and secondary nodes to function. When pfsync is active and properly configured, all nodes will have knowledge of each connection flowing through the cluster. If the master node fails, the backup node will take over and clients will not notice the transition since both nodes knew about the connection beforehand.

First on the primary node and then on the secondary, perform the following:

1. Navigate to **System > High Avail Sync**
2. Check **Synchronize States**
3. Set **Synchronize Interface** to *SYNC*
4. Set **pfsync Synchronize Peer IP** to the other node. Set this to 192.168.13.246 when configuring the primary node, or 192.168.13.245 when configuring the secondary node
5. Click **Save**

Additionally, to make the job on pfsense faster, and easier by using XML-RPC. After enable XML-RPC, the settings from supported areas of the primary node will copy to the secondary and activate if anything on the primary node changes.

On the primary node only, perform the following:

1. Navigate to **System > High Avail Sync**
2. Set **Synchronize Config to IP** to the Sync interface IP address on the secondary node, 192.168.13.246
3. Set **Remote System Username** to admin .
4. Set **Remote System Password** to the admin user account password, and repeat the value in the confirmation box.
5. Check the boxes for each area to synchronize to the secondary node. For this guide, as with most configurations, all boxes are checked. The **Toggle All** button may be used to select all of the options at once, rather than selecting them individually.
6. Click **Save**

Furthermore, with configuration synchronization in place, the CARP Virtual IP addresses need only be added to the primary node and they will be automatically copied to the secondary.

1. Navigate to **Firewall > Virtual IPs** on the primary node to manage CARP VIPs
2. Click Add at the top of the list to create a new VIP.

Type

3. Defines the type of VIP, in this case *CARP*.

Interface

4. Defines the interface upon which the VIP will reside, such as *WAN*

Address(es)

5. The **Address** box is where the IP address values are entered for the VIP. A subnet mask must also be selected and it must match the subnet mask on the interface IP address. For this example, enter 198.168.58.2 and 24 (See Wan IP address Assignments).

Virtual IP Password

6. Sets the password for the CARP VIP. This need only match between the two nodes, which will be handled by synchronization. The password and confirm password box must both be filled in and they must match.

VHID Group

7. Defines the ID for the CARP VIP A common tactic is to make the VHID match the last octet of the IP address, so in this case choose 2

Advertising Frequency

8. determines how often CARP heartbeats are sent.

Base

9. Controls how many whole seconds elapse between Heartbeats, typically *1*. This should match between cluster nodes.

Skew

10. Controls fractions of a second (1/256th increments). A primary node is typically set to 0 or 1, secondary nodes will be 100 or higher. This adjustment is handled automatically by XML-RPC synchronization.

Description

11. Some text to identify the VIP, such as WAN CARP VIP .

    The LAN VIP would be configured similarly except it will be on the *LAN* interface and the address will be 10.0.0.252

The next step will be to configure NAT so that clients on the LAN will use the shared WAN IP as the address.

1. Navigate to **Firewall > NAT**, **Outbound** tab
2. Click to select *Manual Outbound NAT rule generation*
3. Click **Save**

A set of rules will appear that are the equivalent rules to those in place for Automatic Outbound NAT. Adjust the rules for internal subnet sources to work with the CARP IP address instead.

1. Click to the right of the rule to edit
2. Locate the **Translation** section of the page
3. Select the WAN CARP VIP address from the **Address** drop-down
4. Change the Description to mention that this rule will NAT LAN to the WAN CARP VIP address

In order to do configuration synchronization on Squid proxy server & squidGuard by using XML-RPC are need to configure manually.

Squid proxy server

1. Navigate to **Service > Squid proxy server**, **sync** tab
2. Select the sync method for squid on **Enable Sync**
3. Click enable on Replication Targets
4. Choose protocol that we use, in this case we use https and choose port, in this case we use port 443
5. Set **IP** to the Sync interface IP address on the secondary node, 192.168.13.246
6. Set Username, in this case we use Admin
7. Set password

SquidGuard

For SquidGuard would be configured similarly except it will be Navigate to **Service > SquidGuard proxy filter**, **sync** tab

Result

After we have been implemented successful both nodes GUI will look like this.

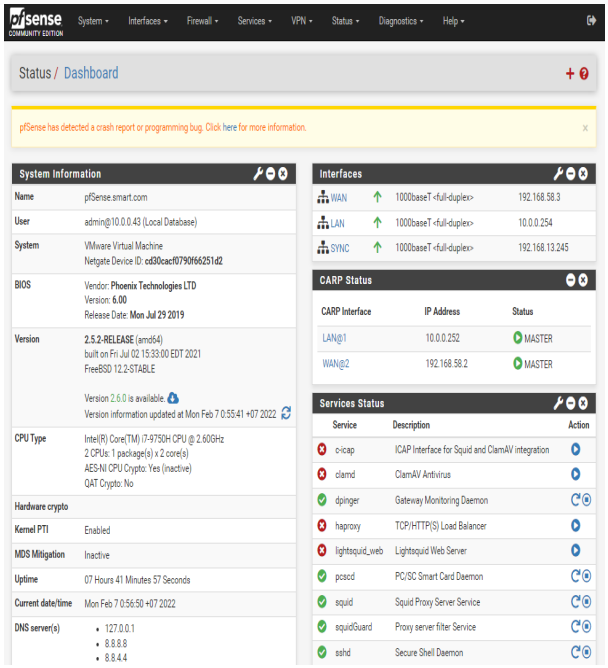


Figure.5 On primary

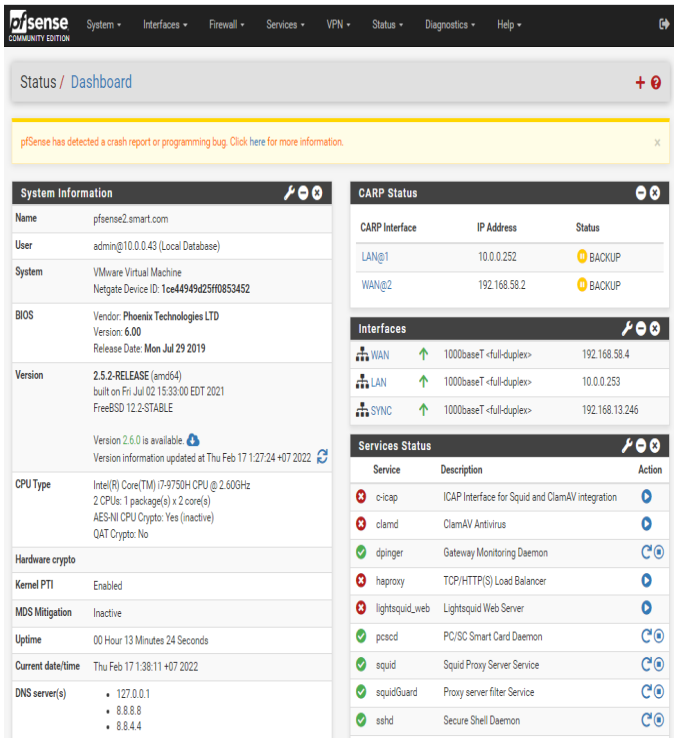


Figure.6 on Secondary

IV. CONCLUSIONS

As a result, after implementing high availability of proxy server on pfsense , it eliminates a single point of failure. When the proxy server went down, another proxy server will take over and the service still available properly. In addition, with enable PfSense XML-RPC config Sync, it makes a job faster and easily, when configured on the primary, it is copied to the secondary proxy server automatically. Last but not least, after implementing high availability of proxy server on pfsense we ensure continuous operation and uptime of at least 99.99% annually.

ACKNOWLEDGMENT

In the accomplishment of this paper successfully, many people have best owned upon me their blessings and the heart pledged support. I would like to express my sincere gratitude to CADT, by organizing the first Student Conference on Digital Technology in 2022, It is the first annual conference on digital technology for recent graduates and bachelor's degree holders. Equally, I would like to show my appreciation to my advisor Mr. Nhem Thayheng who is working so hard to help, guide, encourage, teach and give me powerful support. When I came across an obstacle.

REFERENCES

[1] Fortinet. (2022). Accessed: November 4,2022. [online]. Available: https://www.fortinet.com/resources/cyberglossary/proxy-server

[2] OPNsense. Accessed: November 10, 2022. [online]. Available: https://opnsense.org/about/about-opnsense/

[3] Netgate. (2022). Accessed: Octerber 29,2022. [online]. Available: https://docs.netgate.com/pfsense/en/latest/general/index.html

[4] Netgate. (2022). Accessed: Octerber 20,2022. [online]. Available: https://docs.netgate.com/pfsense/en/latest/highavailability/index.html #carp-overview

[5] Netgate. (2022). Accessed: Octerber 29,2022. [online]. Available:https://docs.netgate.com/pfsense/en/latest/highavailability/ xmlrpc-sync.html

[6] Netgate. (2022). Accessed: Octerber 29,2022. [online]. Available:https://docs.netgate.com/pfsense/en/latest/highavailability/ pfsync.html

Web Server Redundancy Using Nginx

Sopheak Preab, Thayheng Nhem

sopheak.preab@student.cadt.edu.kh, thayheng.nhem@cadt.edu.kh

Cambodia Academic of Digital Technology
Institute of Digital Technology
Department of Telecommunications and Networking

***Abstract*—To be an organization, which plays an essential role in providing web content and some services, requires a backup of applications. Backup today is a lifeline to the future for small, medium, and enterprise organizations alike. That's because data is the world's most valuable resource, and you must protect it. At the same time, all these requirements demand a lot of resources, especially servers are very vital for these requirements in order to either run or host websites, and applications. Furthermore, it is such a big concern for an organization in order to process its operation smoothly since the organization needs to invest a lot of money in servers and need space to store them. Consequently, to respond to the troubles of this project, Docker virtualization technology is suitable to solve the problem to reduce cost. Especially, the container will boost the performance of each virtual machine by spending less time than other virtualization technology platforms with creating, booting, and removing each virtual machine, and it allows a single OS on a machine to host multiple isolated environments, so Docker guarantees to utilize resources in the server as much as possible. All of these will truly respond to and solve the problems.**

***Keywords*— Redundancy, Webserver Software, Virtualization, keepalived.**

I. Introduction

In the present era noticed on our modern world very easy to access any information from the website by anyone, anytime, anywhere if we have internet. But behind the convenience of the web, hidden the complexity of providing the web services. But the main problems often occur on web servers is about availability to manage the data and application to serve any processes. Basically, averaged over a year, web servers are required to be online at 99% within one year that mean they have a 1% down time. Thus, the conventional server would be down for 87 hours and 36 minutes per year or over 3 days [1]. That's why most of the companies or organizations always implement their own data centre contains many servers to operate the working process more efficient and easier to develop which in turn created the number of challenges.

This led to problems such as high cost and poor scalability as each server was both expensive and took up space in the datacentre. By encountering the above problems, professional engineers also developed a new technology called "Virtualization". This technology allows creating the multiple virtual servers and putting it to work efficiently. Anyways although virtualization can create multiple virtual servers, but all those virtual servers are run on a single physical server, so when that physical server down the processes are dead.

We need the solution that web service still run well even though the main server down to ensure the web service work efficiently and high availability.

There are various kinds of web servers across the internet today. Below is the amount of website increasing from 1995 to 2022 that was surveyed for Netcraft.

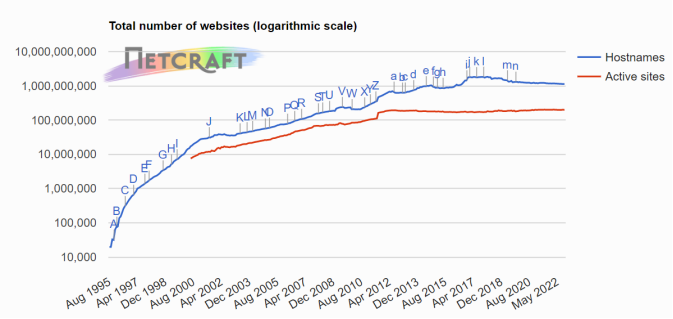


Figure 1: Amount of website increasing

Source: Netcraft(https://news.netcraft.com/)

II. Literature Review

In the past few years, the world wide web has experienced phenomenal growth. Not only are millions browsing the web, but hundreds of new websites are added each day [2]. Yet, despite the increasing number of Web servers in use little is definitively known about their performance characteristics [3].

It is widely accepted that increasing the level of redundancy in a system generally improves service availability and system dependability [4].

III. Methodology

A. Main Concept

The organization wants to build the secondary web server for hosting the website and make it fail-over architecture, when the primary server down it has fail-over so the secondary server can work instead the primary automatically. In this, we are going to find out the experiences of virtualization and software for web serving platform, which is suitable for our infrastructure in order to improve performance in each virtual machine and save resources in server. Therefore, Docker is the containerize virtualization platform responding to the purpose of this project since Docker container will boost performance of each virtual machine by

spending less time than before with creating, booting, and removing each virtual machine easily. About the software for web serving, Nginx is the best and fastest open-source technology suitable in our project which is the latest technology most supported today. Nginx beyond the web serving, it also reverses proxying, caching, load balancing, media streaming, and more. It started out as a web server designed for maximum performance and stability. In addition to its HTTP server capabilities. Therefore, these technologies will work together to improve the performance and ability of the service.

B. Flow of Diagram

In order to demonstrate about our research, we also have a topology to represent about the process of the topic.

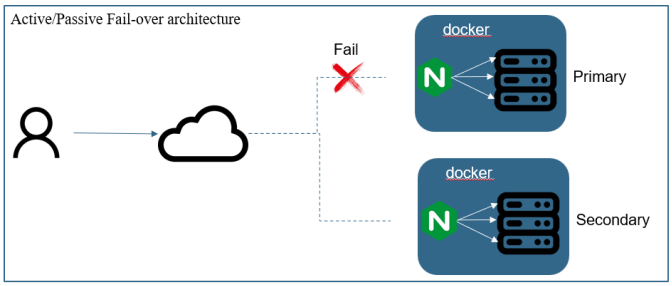


Figure 2: Topology

As the figure 1 show above, the infrastructure wants to build two servers and each server implement with the same thing. Each of them uses Docker technology and using Nginx for serve the web serving. Both have redundant with each other. Commonly when the user accesses the web from the internet, the primary server will serve the service and respond to the request. Whenever the primary server fails, the secondary will automatically work instead immediately and perform the same thing.

C. Redundancy

Redundant system will provide failover or load balancing to protect a live system in the event of an unexpected failure. In the case of power, mechanical, or software failure, a redundant system will have a duplicate component or platform to fall back to [5]. To build web server redundancy, the first thing to do is building a web server. The importance thing to build a web server is select the best server product. Many kinds of server products exist in the market, so before choosing the one to implement please think about the infrastructure matched, price, quality, and ability of that server. After selected the right server, the next step is implemented it as web server.

D. Docker

To reduce the cost and space investing on server. The very suitable solution is virtualization technology. Docker is the selected. As it is an open-source platform that run applications and make the process easier to develop, distribute. The application that are built in the docker are packaged with all the supporting dependencies into a standard form call a container. These containers keep running in an isolated way on top of other operating system's kernel [6]. Comparing container with virtual machine.

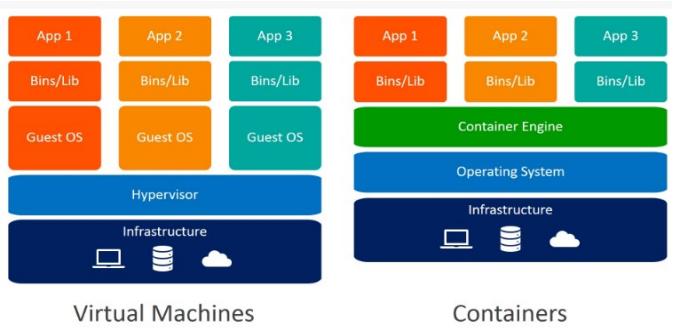


Figure 3: Virtual machine vs Container

Another to know is docker-compose. For applications depending on several services, orchestrating all the containers to start up, communicate, and shut down together can quickly become unwieldy. Docker Compose is a tool that allows you to run multi-container application environments based on definitions set in a YAML file. It uses service definitions to build fully 25 customizable environments with multiple containers that can share networks and data volumes. [7].

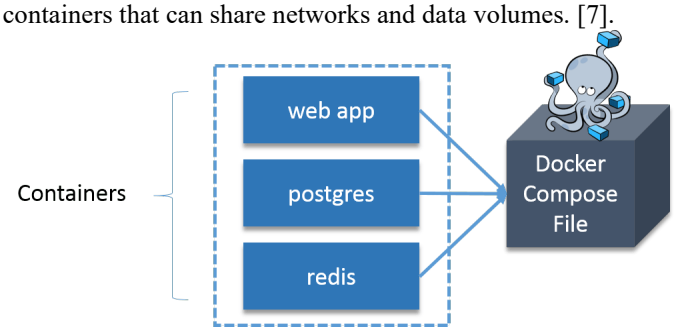


Figure 4: Docker compose

E. Nginx

There are many webs server software are uses in the market today. From day to day and year to year, developers' effort to develop new technologies to fit their needed. And each of them has their own unique and benefits. Below is the diagram to represent about market share of all sites from 1995 till 2022 that was surveyed from Netcraft.

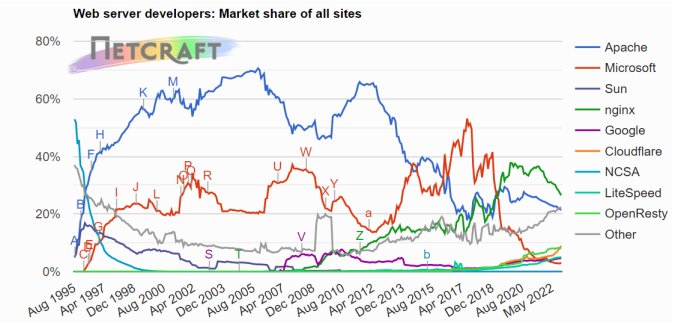


Figure 5: Market share all sites Source: Netcraft(https://news.netcraft.com/)

According to figure 3 above we can see that nowadays Nginx is very popular and top using in the market although its significant losses in its number of sites and domains. However, Nginx still holds its strong lead as the most widely used web server software, with a share of 26.51% sites. And Apache is the second largest number of sites.

Nginx beyond the web serving. It can serve HTTP and HTTPS, proxy, load balancer, etc. Nginx offer low memory usage and high concurrency. Rather than creating new processes for each web request, Nginx uses an asynchronous, event-driven approach where requests handled in a single thread. With Nginx, one master process can control multiple worker processes. The master maintains the worker processes, while the workers do the actual processing. Because Nginx is asynchronous, each request can be executed by the worker concurrently without blocking other requests [8].

IV. IMPLEMENTATION

In order to make the web server have redundant, must have at least two servers to handle the process. To make easier understanding, figure 5 below will display the detail process of the diagram.

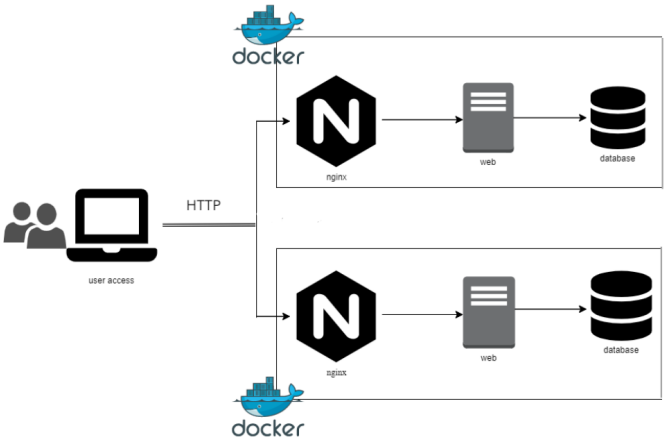


Figure 6: diagram process

Refer to figure 6, the implementation divided into 3 steps:

1) Implement Docker and Docker-compose: the first thing to do is using command line to download docker. To ensure to get the original package, should download from Docker repository with the specific version or latest version. To do that, we will add a new package source, add the GPG key from Docker to ensure the download are valid first, and then install the package and update the server again. After install docker successfully download docker compose and install it.

2) Deploy Container: Here must create three containers to process the service. One is container to work as the web server, another one is the container to work as the web application and the last one is work as database. All containers can build and up at the same time by using docker compose. First, create network for container to communicate with each other. Next create docker file. After that define services in docker compose script. First, we must pull the image for the container then assign the container name. Define working directory, defines the volume then assign the specific IP for each container. when it finishes, build and run applications with docker compose command. Finally update the application.

3) *Customize nginx script:* as shown in figure 6 nginx web server must access to the web. So, we must create the path from nginx to web that we must assign the port to work for nginx. Determine the index and finally add the location of the web.

4) *Redundant Configuration:* using command line to install keepalived technology then download the tarball for keepalived. Next install script tool and setting the init script. In the configuration must define the state, interface, priority (priority in master server is lower than priority in backup server), peer address and assign specific ip for each. At the end, start the service.

IV. CONCLUSION

After implement the redundancy for the web server, the service performance work very well. When the primary server down, the users don't even know that the server has problems since the service work properly. This because of the redundancy process that let the secondary server provide the content as usual. Anyway, our backend can be more secure by using Nginx, it beyond the web server.

ACKNOWLEDGMENT

The successful and outcome of this research paper were possible by support from many people. It required a lot of effort from everyone involved in this research, and I would like to thank them. I appreciate and thank to CADT that organize the first student conference on digital technology 2022 that is the first annual conference on digital technology for bachelor's degrees and fresh graduate students. Especially, I would like to show my highest respect to my advisor Mr. Nhem Thayheng who always support and guide from the beginning till the end of the research.

REFERENCES

[1]. J. Kyrnin, "What Is Uptime in Web Hosting," ThoughtCo,pp. thoughtco.com/uptime-in-web-hosting-3467355, Sep. 3, 2021.
[2]. D. A. A. S. Yih Huang, "Closing Cluster attack windows through server redundancy and rotations," conference paper, p. 1, 2006.
[3]. J. a. C. S. iederspan, "Planning and managing web sites on the macintosh," Addison-Wesley, 1996.
[4]. L. P. Slothouber, "A Model of Web Server Performance," 1995.
[5]. D. DeJonghe, "Complete NGINX Cookbook," Part I: Load Balancing and HTTP Caching, pp. 11-13, 2017.
[6]. Atlantic, "Overview of redundant server," 5 Dec 2022. [Online]. Available:https://www.atlantic.net/resources/documents/whitepapers/Overview-of-Redundant-Systems-Atlantic-Net-Whitepaper.pdf.
[7]. Boettiger, "An introduction to Docker for," ACM SIGOPS Operating Systems, pp. 71-79, 2015.
[8]. J. Cook, "Docker Compose," in Docker for Data Science, California, 2017, pp. 179 – 180.

ABOUT THE CONFERENCE

The 1st Student Conference on Digital Technology 2022 (1st SCDT) is the first annual conference on digital technology for undergraduate and fresh graduate students. The conference highlights the research in areas such as digital platforms, data monetization, digital materials, wireless and satellite communications, human-computer interactions, Machine learning and data science in process automation, formal models for design reuse decisions, digital innovation, and especially, Internet of Thing (IoT), Artificial Intelligence, Blockchain and Cloud (IABC).

CADT

បណ្ឌិត្យសភាបច្ចេកវិទ្យាឌីជីថលកម្ពុជា
Cambodia Academy of Digital Technology

CONTACT INFORMATION
H/P: (+855)10 340 000
Email: pr@cadt.edu.kh
Website: www.cadt.edu.kh
Address: National Road 6A, Kthor, Prek Leap Chroy Changvar, Phnom Penh, Cambodia