

Classifying and Blacklisting Phishing Webpage Using NCD and Image Hotspot Security

Ms.M.Abinaya

UG Student Computer Science and Engineering, Sri Ramakrishna
College of Engineering, Perambalur Tamilnadu, India.

Ms.R.Mohanapriya

UG Student, Computer Science and Engineering
Sri Ramakrishna College of Engineering, Perambalur, Tamilnadu, India.

Ms.S.Ranjitha

UG Student, Computer Science and Engineering
Sri Ramakrishna College of Engineering, Perambalur. Tamilnadu, India.

Ms.G.Shiraginisha

UG Student, Computer Science and Engineering
Sri Ramakrishna College of Engineering, Perambalur. Tamilnadu, India.

****Ashish Jjoshi ,Sonukumar,Yang Jinbo,Mohammed Ali Sakidin,Jun
Nishikaw,Abey Abraham,Xeumincheng,Hui Tian,Tian wang,Gamze
Akman,Ying Li,Ziping du,Ning Zhang,Pradeep Atrey.**

Abstract

In this project, we propose a feature-free method for detecting phishing websites using the Normalized Compression Distance (NCD), a parameter-free similarity measure which computes the similarity of two websites by compressing them, thus eliminating the need to perform any feature extraction. It also removes any dependence on a specific set of website features. This method examines the HTML of webpages and computes their similarity with known phishing websites, in order to classify them. We use the Furthest Point First algorithm to perform phishing prototype extractions, in order to select instances that are representative of a cluster of phishing webpages. We also introduce the use of an incremental learning algorithm as a framework for continuous and adaptive detection without extracting new features when concept drift occurs.

Date of Submission: 29-05-2022

Date of acceptance: 10-06-2022

I. INTRODUCTION

An additional feature we added into this project Image hotspots. It can be useful for creating info graphics fast and simple. Use any image and enrich it with points of interest and in-depth information about the details depicted. The user is activated by interacting with the image. When the user clicks inside the image the browser will append the X and Y coordinates (relative to the upper-left corner of the image) to the anchor URL as a query string and will access the resulting URL. It allows you to create different gestures on any image of your choice and use those gestures as your password. The gesture can be any combination of circles, pixels, and taps. For example if the picture you chose was of a face your picture password could be a tap on each eye and then a circle around the mouth.

II. Existing System

There is significant research focusing on phishing detection. Some studies focus on the use of blacklists and whitelists in anti-phishing systems. Blacklist-based methods keep a list of domain names or links to known phishing websites and alert users if they try to visit those sites. However, phishing websites are highly dynamic, and the average lifetime of a phishing webpage is only a few hours. In many cases, zero-hour phishing attacks easily bypass blacklist-based methods. Meanwhile, whitelist-based approaches allow users to browse only those

webpages that are deemed safe, which is often impractical. Other studies in phishing detection use similarity-based methods to measure similarities between websites by analyzing a website's textual content or screenshot. Text similarity-based methods analyze the semantics of the textual content of emails and webpages to decide whether they classify as a phishing attempt. This method is likely to fail in the future with the increased use of code obfuscation techniques.

Disadvantages:

- (a) The app can reach large numbers of users and their friends to spread spam,
- (b) The app can obtain users' personal information such as email address, home town, and gender, and
- (c) The app can "re-produce" by making other malicious apps popular.

III. Proposed System

We introduce a systematic method to perform website similarity measurements for detecting similar phishing websites using Normalized Compression Distance (NCD). We provide an analysis on the similarities and differences between phishing and legitimate website contents and visual appearances, and how content-based methods would effectively detect phishing attacks better than visual-based methods. We propose PhishSim as a tool to effectively detect slightly modified or near-similar phishing websites using prototype-based learning algorithms and the Normalized Compression Distance, which is a parameter-free and application independent distance metric to measure similarities between websites' HTML content. This tool works by measuring the pairwise similarity between websites in the dataset, clustering these websites, and performing phishing classifications based on whether a website is grouped in the same cluster with a known phishing website. Pixel is a sample of an original image; more samples typically provide more accurate representations of the original. The intensity of each pixel is variable. In color image systems, a color is typically represented by three or four component intensities such as red, green, and blue, or cyan, magenta, yellow, and black. The term pixel is used to refer to a single scalar element of a multi-component representation

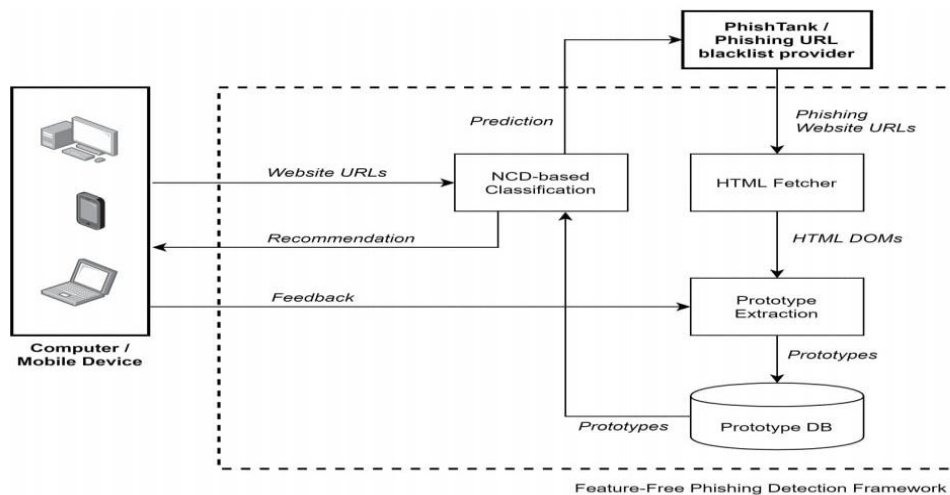
Advantages

- (a) Personal information mostly secured for phishing webpages.
- (b) Image hotspot security prevented to hackers are don't misuse by their information.
- (c) Secured authentication including their data privacy and security.

IV. LITERATURE SURVEY

S.NO	PAPER TITLE WITH YEAR PUBLICATION	AUTHOR	METHODOLOGY	ADVANTAGES	DISADVANTAGES
1.	A More Multifactor Phishing Web scheme based on Graphical (Jan 2018).	Ashish Josh, Sonukumar	We generally use knowledge based technique. Eg. textual password	Deploy targeted anti-phishing solution.	The current authentication system suffer from many drawbacks.
2.	Aiding Phishing Website detection (Apr 2001).	Yang Jinbo, S	Password based authenticated protocol.	Protect valuable corporate and personal data.	Loss of productivity, Loss of company value.
3.	Webpage simulation on phishing webpage signcryption (Aug 2018).	Mohammed Ali Sakidin	Nowadays mobile banking becomes a popular tool which consumer can conduct financial.	Financial transaction, transferring funds and other payments.	"Re-produce" by making malicious apps popular.
4.	Quantum secure authentication system using adaptive optics (May 2019).	Jun Nishikaw	The authentication technique in prevent replicating a physical key and digital emulating by an unauthorized access.	Quantum secure authentication is a novel authentication.	Key and Digital emulating by an authorized access.
5.	NCD for measuring Phishing website similarity (Sep 2020)	Abey Abraham	Such a credit card or NCD for measuring phishing website similarity debit card.	In today word is a wide range term for theft and fraud committed using or involving a payments card.	Personal information such as email address home town and gender.

V. SYSTEM ARCHITECTURE



VI. MODULES

1. Phishing Website Classification
2. Prototype-based Learning
3. Phishing Prototype Database Update
4. URL Validation
5. URL Blocking
6. Image Hotspot Security

VII. MODULES DESCRIPTION

Phishing Website Classification:

- NCD –based similarities website and phishing prototypes.
- Prototype database or blacklist provider, and redirected to warning page.

Prototype-based Learning:

- Two websites perform clustering to divide HTML contents into number of group.
- Small number of pairwise NCD values.

Phishing Prototype Database Update:

- Prototype database periodically receiving data from phishing blacklist.
- To extract representative prototypes from new data.

URL Validation:

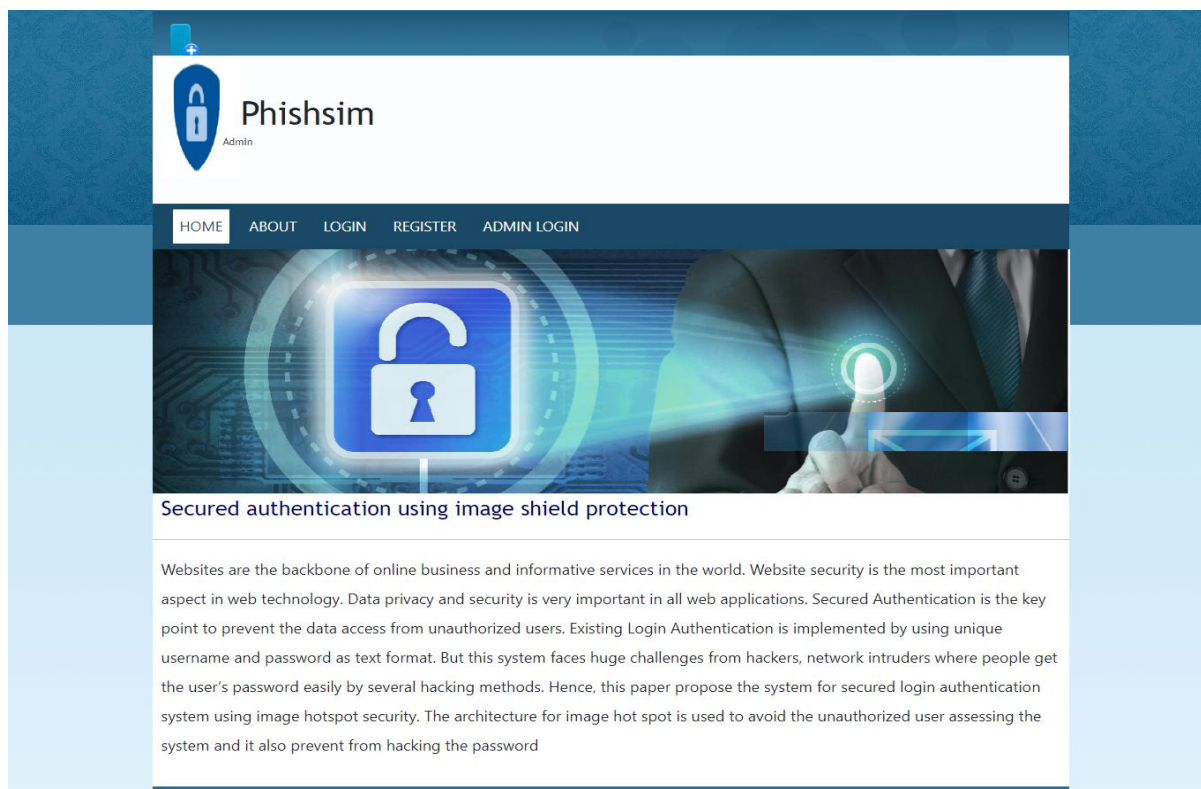
- Browser identified url is valid or not
- The webpage detecting whether executing the same request server or not.

URL Blocking

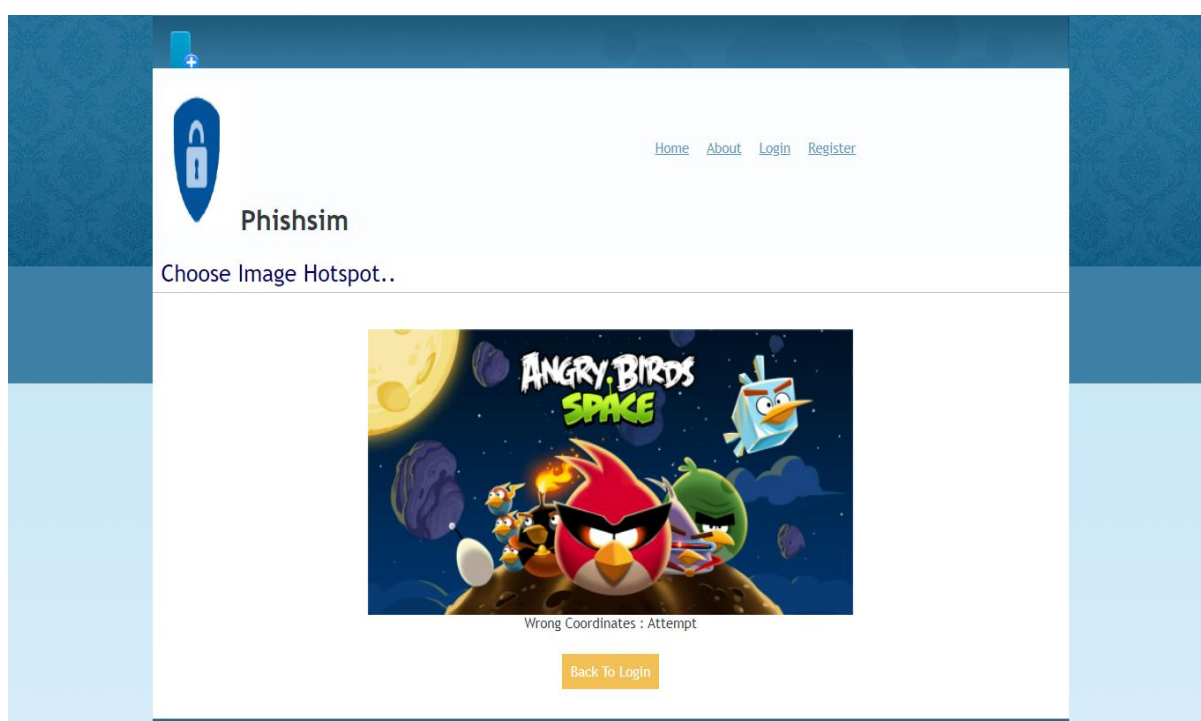
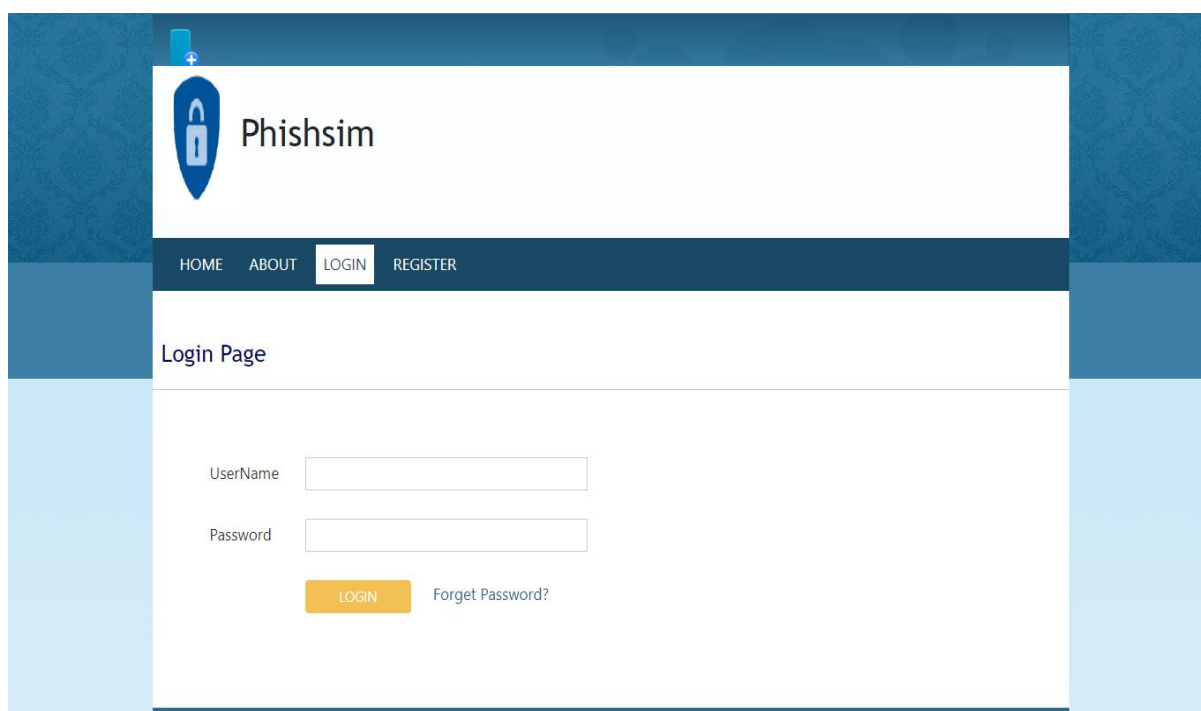
- Users can enter the url in the developed browser which can be identified that the entered url is valid or not. For the url entered by the user, the browser helps in displaying the web page and the entire source code have been crawled for detecting whether it is executing within the same requested server or not.

Image Hotspot Security

- The architecture for image hot spot is used to avoid the unauthorized user assessing the system and it also prevent from hacking the password.
- High probability of finding the password, proposed system with one hotspot password is designed.



The screenshot shows the Phishsim website's registration form. At the top left is a logo featuring a blue shield with a white padlock and the word 'Admin' below it. To the right of the logo is the text 'Phishsim'. Below the logo and text is a navigation bar with links: HOME, ABOUT, REGISTER, and LOGIN. The main content area features the heading 'Sign Up... Step-1'. Below the heading is a registration form with the following fields: Name, Email, Mobile No, Address, Date of Birth (with a date picker), Gender (with a dropdown menu), UserName, Password, Security Question (with a dropdown menu), and Answer. At the bottom of the form is a yellow button labeled 'SUBMIT'.



VIII.CONCLUSION

This method examines the HTML source codes of webpages and computes their similarity with known phishing websites. First algorithm to perform phishing prototype extractions, in order to select instances that are representative of a cluster of phishing webpages. We also introduce the use of an incremental learning algorithm as a framework for continuous and adaptive detection without reperforming new feature extraction when concept drift occurs.

REFERENCES

- [1]. "Google Custom Search," <https://developers.google.com/custom-search/>, accessed: 2019-03-04.
- [2]. "Phishers' Favorites: After Five Quarters, Microsoft is Unseated by PayPal," <https://www.vadsecure.com/en/phishers-favorites-q3-2019/>, accessed: 2020-04-12.

- [3]. "Phishers' Favorites: It's Lonely at the Top: Microsoft Remains the #1 Impersonated Brand in Phishing Attacks," <https://www.vadese.com/en/phishers-favorites-q1-2019/>, accessed: 2020-01-28.
- [4]. "Phishers' Favorites: Microsoft Holds Onto the #1 Spot but Facebook Phishing Surges," <https://www.vadese.com/en/phishers-favorites-q22019/>, accessed: 2020-04-12.
- [5]. "Phishers' Favorites: PayPal Leads, Note Phishing Increases, and Smaller Banks Become Bigger Targets," <https://www.vadese.com/en/phishers-favorites-q4-2019/>, accessed: 2020-04-12.
- [6]. "PhishTank: An Anti-Phishing Site," <https://www.phishtank.com/>, accessed: 2018-10-1.
- [7]. "The Chromium Projects: Chromium," <https://www.chromium.org/Home>, accessed: 2020-07-27.
- [8]. "Using Custom Elements," https://developer.mozilla.org/enUS/docs/Web/Web_Components/Using_custom_elements, accessed: 2021-12-17.
- [9]. Aleroud and L. Zhou, "Phishing Environments, Techniques, and Countermeasures: A Survey," *Computers & Security*, vol. 68, pp. 160–196, 2017.
- [10]. APWG, APWG Phishing Trends Reports. Anti Phishing Working Group, 2016.

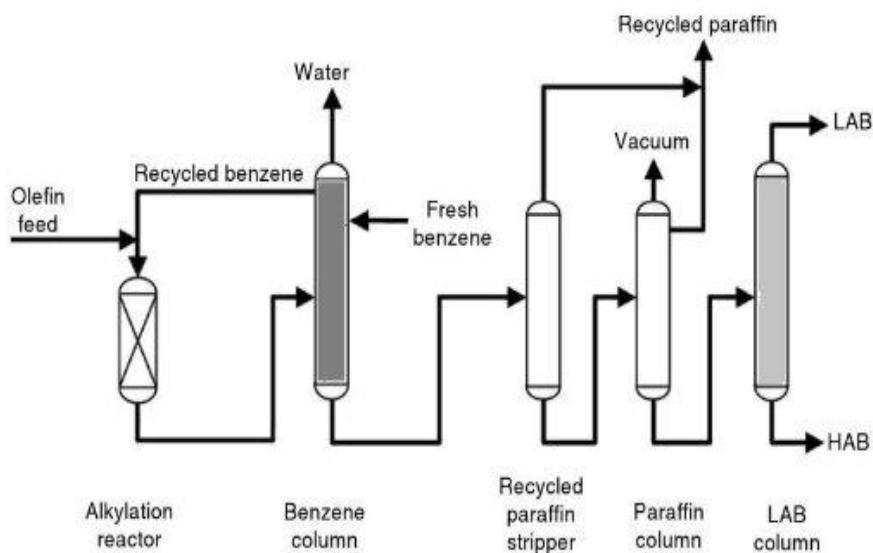


Figure1: Linear alkylbenzene flow diagram (UOP 2009[2])

1.2 THE SIMULATION

1.2.1 The Linear alkylbenzene (rerun column) was simulated at steady state using a distillation column template of ASPEN HYSYS 8.8. Peng Robinson was selected as the fluid package.

Figure 2:represent the simulated flow diagram for the linear alkylbenzene (rerun) column

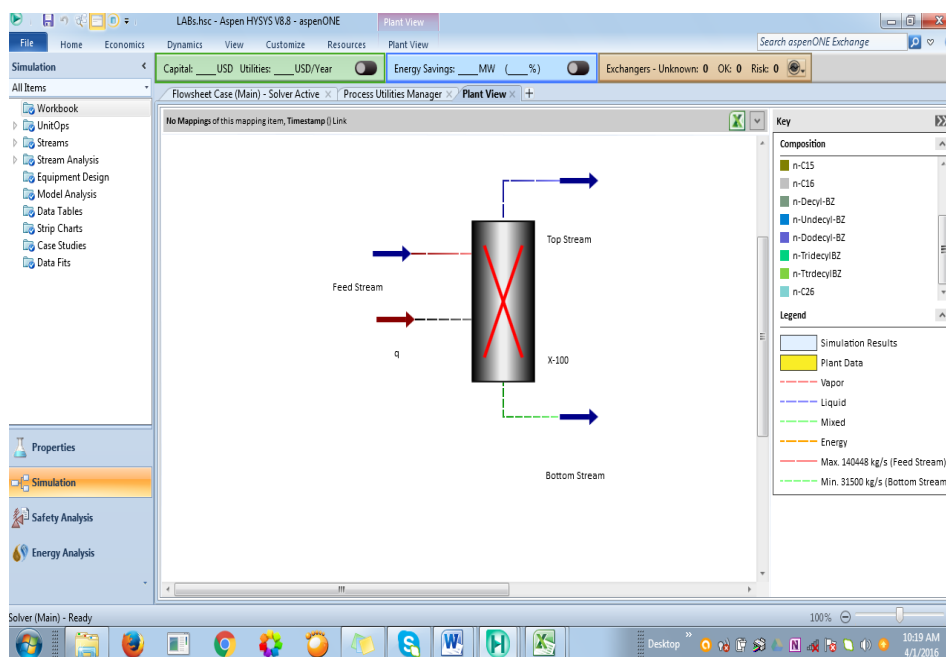


Figure 2:Aspen HYSYS linear alkylbenzene Column Plant View

1.2.2 Rerun column Design specifications

Table 1 represents the rerun column design specification.

Table 1: Rerun Column, Specification.

	Specification
Striping Column diameter	2800mm
Rectification Section diameter	5600mm
Tray spacing	600mm
Number of tray holes	1942
Hole diameter	13mm
Number of Trays above feed	15 trays, 16 stages (with condenser)
Number of Trays below feed	21 trays, 22 stages (with re-boiler)
Q_c (Condenser heat duty)	31212 MJ/hr
Q_r (Re-boiler heat duty)	22363 MJ/h

1.2.3 Feed Components and composition

Table 2 indicates the LAB column feed stream components the composition in wt. % fraction.

Table 2: Feed Components used and their Composition in weight % fraction.

Components	Chemical Formula	Composition (wt. % fraction)
1. N-Decane	$nC_{10}H_{22}$	0.0003
2. N-Undecane	$nC_{11}H_{24}$	0.0033
3. N-Dodecane	$nC_{12}H_{26}$	0.0046
4. N-Tridecane	$nC_{13}H_{28}$	0.0058
5. N-Tetradecane	$nC_{14}H_{30}$	0.0065
6. N-Pentadecane	$nC_{15}H_{32}$	0.1086
7. N-Hexadecane	$nC_{16}H_{34}$	0.0166
8. Decylbenzene	$nC_{169}H_{26}$	0.1439
9. N-undecylbenzene	$nC_{17}H_{28}$	0.1775
10. N-dodecylbenzene	$nC_{18}H_{30}$	0.2032
11. N- tridecylbenzene	$nC_{19}H_{32}$	0.1626
12. N-tetradecylbenzene	$n-C_{20}H_{34}$	0.1071
13. Heavy Alkylates	$n-C_{26}H_{54}$	0.0599

2.2.4 Rerun column Operating Conditions

The operating conditions of the LAB (rerun) column is as represented in table 3

Table 3 Operating Conditions used in this Simulation are as stated below.

	Feed	Top Stream (Distillate)	Bottom
Temperature °C	178	93	232
Pressure Kpa	200	9.0	20
Enthalpy MJ/h	-95850	-98130	-125900
Mass Flow-rates Kg/h	140400	108900	14290

III. RESULT AND DISCUSSION

The results obtained are as discussed below

1.3.1 LAB Average Weight fraction.

(i) At Top stream Operating temperatures $\Delta T=20^\circ\text{C}$

The average weight fraction of n-decylbenzene, n-undecylbenzene and n-dodecylbenzene increased at the top stream operating temperature of $280^\circ\text{C} - 300^\circ\text{C}$ and decreased at $320^\circ\text{C} - 360^\circ\text{C}$. While n-tridecylbenzene and n-tetradecylbenzene decreased at $280^\circ\text{C} - 300^\circ\text{C}$ and increased at $320^\circ\text{C} - 360^\circ\text{C}$.

Table 4: Average Weight fraction of LAB at Top Stream Operating Conditions $\Delta T=20^\circ\text{C}$.

Linear Alkylbenzenes (LAB)	Operating Temperatures(°C)				
	280	300	320	340	360
n-decylbenzene	0.2792	0.2984	0.2849	0.2791	0.2740
n-undecylbenzene	0.2177	0.2800	0.2740	0.2714	0.2694
n-dodecylbenzene	0.2842	0.2762	0.2807	0.2820	0.2832
n-tridecylbenzene	0.1658	0.1566	0.1638	0.1675	0.1705
n-tetradecylbenzene	0.0885	0.0832	0.0889	0.0927	0.0950

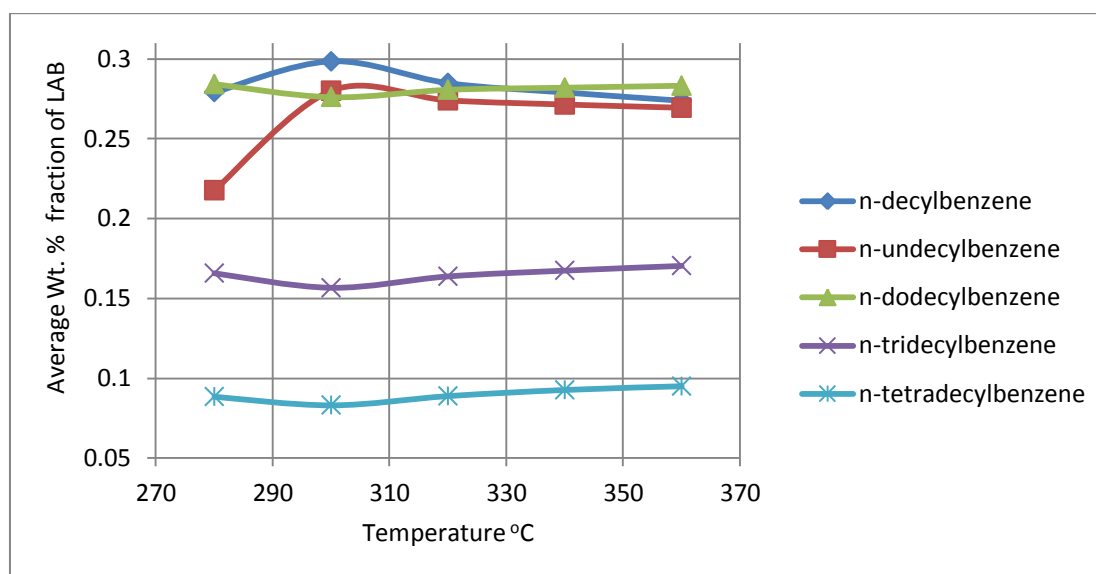


Figure 3: Average Weight fraction of LAB at Bottom Stream Operating Conditions

(ii) At Bottom stream Operating temperatures $\Delta T=20^{\circ}\text{C}$

The average weight fraction of n-decylbenzene, n-undecylbenzene and n-dodecylbenzene increased at the bottom stream operating temperature of $280^{\circ}\text{C} - 320^{\circ}\text{C}$ and decreased at $340^{\circ}\text{C} - 360^{\circ}\text{C}$. While that of n-tridecylbenzene and n-tetradecylbenzene remains constant. Table 4 and Fig.4

 Table 4: Average Weight fraction of LAB at Bottom Stream Operating Conditions $\Delta T=20^{\circ}\text{C}$.

Linear Alkylbenzenes	Operating Temperatures($^{\circ}\text{C}$)				
	280	300	320	340	360
n-decylbenzene	0.2390	0.2393	0.2404	0.2419	0.2401
n-undecylbenzene	0.2558	0.2561	0.2568	0.2569	0.2568
n-dodecylbenzene	0.2797	0.2808	0.2831	0.2829	0.2837
n-tridecylbenzene	0.1897	0.1897	0.1897	0.1897	0.1897
n-tetradecylbenzene	0.1113	0.1113	0.1113	0.1113	0.1113

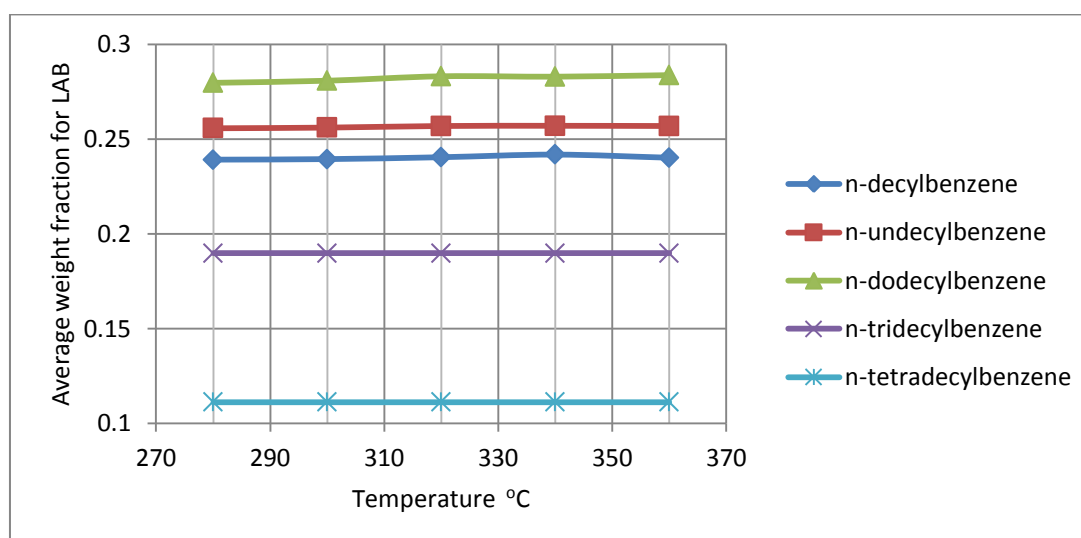


Figure 4: Average Weight fraction of LAB at Bottom Stream Operating Conditions

4.2.4 LAB Average wt. % fraction Yield

The calculated LABs percentage Yield from the average LAB weight fraction is as shown in Table 4.3

The calculated percentage yield of average LAB wt. % fraction indicated a higher percentage yield at the top and bottom stream temperature of 280°C. At this temperature the obtained yield of top and bottom streams are 92.2% and 95.3 %. Table 5 and Figure

Table 5: LAB Average wt. % fraction Percentage Yield

Operating *Temperatures (°C)	Percentage Yield of Average LAB wt.% at various Top Stream operating condition	Percentage Yield of Average LAB wt.% at various Bottom Stream operating condition
280	92.2	95.3
300	87.6	94.8
320	87.5	94.2
340	87.6	93.8
360	87.5	93.3

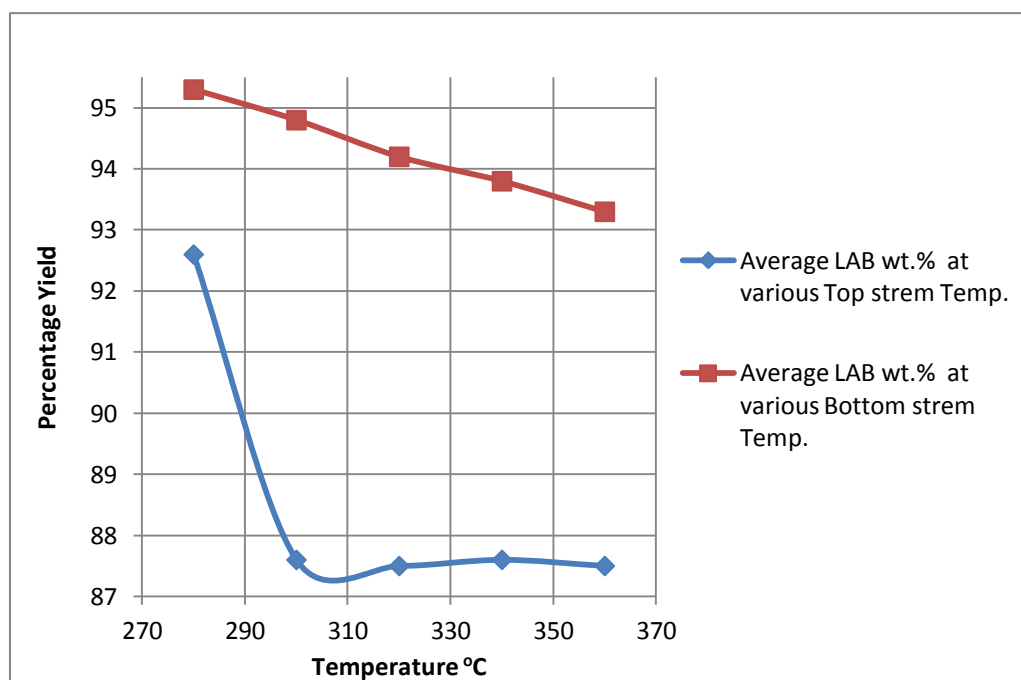


Figure 5: LAB Average wt. % fraction Percentage Yield

5 Percentage Yield of LAB in the distillate at $T_{\text{Bottom}} = 280^{\circ}\text{C}$.

- The bottom stream operating temperature of 280 °C has the highest average percentage yield of linear alkylbenzenes (LABs). The percentage yield of the linear alkylbenzene was calculated by keeping the operating temperature at 280 °C and varying the operating pressure at 17Kpa, 42Kpa, 67Kpa, 92Kpa and 115Kpa. The highest yield obtained is 99.4% which is at 115Kpa. This is as shown in Table 6. And figure 6

Table 6: Percentage Yield of LAB in the distillate.

Pressure Kpa	Percentage Yield Of LAB in the Distillate
17	89.1%
42	95.4%
67	97.9%
92	98.97%
115	99.4%

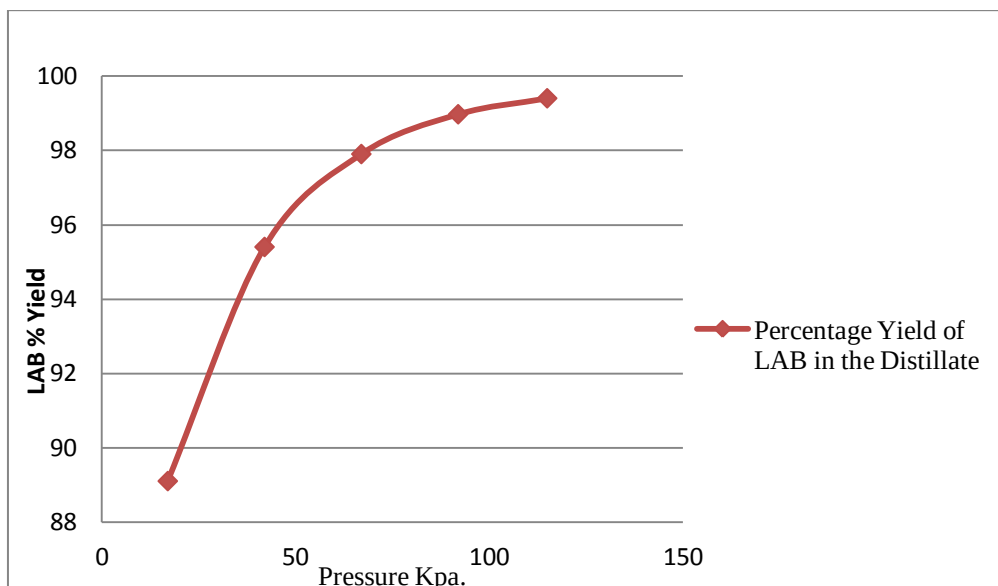


Fig.6:LAB % Yield at $T_{\text{Bottom}} = 280^{\circ}\text{C}$.

IV. CONCLUSION

It was observed that the rerun column bottom stream temperature has greater effect on the linear alkylbenzene yield than the temperature variation of the top stream. At higher temperature of both streams, lower percentage yield of average wt. % of linear alkylbenzene was obtained with that of the top stream being the lowest at 87.5% as against 93.3% for the bottom stream. The highest linear alkylbenzene yield of 99.4% was recorded at bottom stream temperature of 280°C and pressure of 115Kpa.

REFERENCES

- [1]. Abiola K. "Strategy for Development of the Petrochemicals Industry in Nigeria" A paper submitted to the Department of chemical Engineering, University of Lagos.
- [2]. Aderogba, K. A. (2011) "Significance of Kaduna River to Kaduna Refining and Petrochemicals Complex" African Journals, Vol. 5 (5), Serial No. 2 Pp.83-98.
- [3]. Ahmad Daaboul. "LAB Project - Environmental Impact Assessment" Section 3 pp. 4 – 42
- [4]. Irena O. D, Dolganov I.M and Ivanskina E.N (2001) "Development of Computer Modeling System as a Tool for Improvement of Linear Alkylbenzene Production" J. Petroleum and Coal. Vol. 53, No.4, Pp. 244 – 250
- [5]. Sadal O.I., Marwa S.M., Wala T.A. (2012) "Linear Alkylbenzene Production from Kerosene" Seminar presented to the Department of Chemical engineering University of Khartoum.
- [6]. Thaer, A. A. (2010), "Process Simulation Analysis of HF Stripping Column Using HYSYS Process Simulator" J. of Engineering Sciences /Vol.17/No.2, pp.87 – 96.
- [7]. UOP (1990) "Linear Detergent Alkylation Unit, General Operating Manual" pp. 1 – 610.
- [8]. UOP (2004) "Linear Detergent Alkylation Unit, General Operating Manual" Pp. 1 – 112.
- [9]. Xiaoming J, Gang Rong and Shuqing Wang (2003), "Modelling and Advanced Process Control (APC) For Distillation Columns of Linear Alkylbenzene Plant Key Lab of Industrial Control Technology, Institute of Advanced Process Control, Zhejiang University, pp1-6.