Journal of Advanced Zoology

ISSN: 0253-7214
*Volume **44** Issue **S-6 Year 2023** Page **955:960***

Recognizing phishing site using Machine Learning- A Comparative Approach using MultinomialNB & Logistic Regression

Supreeth S., Abhishek Nigam, Akansh Srivastava, Aniket Singh, Ashish Kumar Behera

School of Computer Science and Engineering, REVA University, Bengaluru
supreeth.s@reva.edu.in
abhiaryan972@gmail.com
akanshsri20@gmail.com
aniketsingh.578@gmail.com

**Corresponding author's E-mail: supreeth.s@reva.edu.in*

Article History

Received: 06 June 2023
Revised: 05 Sept 2023
Accepted: 30 Nov 2023

Abstract

Phishing is a method of trying to collect personal information like login credentials or credit card information using deceptive e-mails or websites. Phishing sites are made to hoodwink clueless clients into intuition they are on an authentic site. The lawbreakers will invest a great deal of energy causing the site to appear as valid as could really be expected and numerous locales will show up practically undefined from the genuine article. This paper proposes a methodology to detect boycotted URLs using machine learning algorithms so that people can be frightened while examining or getting to a particular site. In this project we have using machine learning algorithms such MultinomialNB and Logistic Regression. We used distinctive data and text pre-processing techniques to improve precision and accuracy. An app is developed as the end product of this research work.

Keywords: *Phishing, URLs, Machine Learning, NLP*

1. Introduction

Phishing can be depicted as replicating a generous site to swindle clients by using their real factors containing usernames, passwords[15], accounts numbers, public security numbers, and so forth Phishing cheats may be the broadest cybercrime applied today. There are inestimable spaces where phishing assaults can happen like online segment place, webmail, and monetary reason, record working with or cloud limit[16], and different others. The webmail and online segment district had been tormented through phishing more than in some novel industry region. Phishing ought to be appropriate through URL phishing tricks and lance phishing from now on clients need to recall the results and should at this point don't pass on their 100% trust in standard affirmation applications. Machine learning and NLP are one of the productive strategies to choose to phish as it disposes of risks of existing methods. The destination that's the most vital factor within the proposed challenge is to affirm the legitimacy of the web page via catching boycotted URLs. to tell the patron on access boycotted websites through spring up whilst they are trying to get right of entry to and to tell the customer on boycotted websites thru URLs whilst they're attempting to get to. This proposed task will allow the director to feature boycotted URLs to equipped customers all through their request. Right now, individuals achieve maximum online business, transferring cash, charge instalments for instance all of the matters are completed making use of sites or applications. In this way, coming across web page phishing is a huge big thing in our regular existence. It is probably the maximum seasoned type of cyber-attacks, going back to the 90s, it is as but one of the most across-the-board and malicious, with phishing messages and approaches getting regularly superior. Assailants commit notably extra power to deceiving those casualties, who have been selected because the potential prizes are very excessive.

I. Literature survey

The emerging headway industry which fundamentally impacts the current security issues has given a non-ease of the brain to some business and home clients. Events that abuse human inadequacies have been on the upsurge recently. In [1] has zeroed in on a phishing identification framework to recognize boycotted URLs otherwise called phishing sites so people can be alarmed while perusing or getting to

a specific site. In [2] utilized a blend of the Naive Bayes and Decision tree estimations have been assembled using the typical example of Machine Learning showing by recognizing the features of polluted messages by phishing. In [3] performed ELM got from various 30 principal parts which are arranged utilizing the ML approach and utilized 3 different ways for the location of site phishing. In [4] discussed techniques accessible to recognize phishing sites, A similar investigation of the being used enemy of phishing instruments was refined and their impediments were recognized, examined the URL-based highlights utilized in the past to improve their definitions according to the current situation which is our significant commitment. In [5] used major visual highlights of a website page's appearance as the premise of identifying page similitudes and they proposed a framework that utilizes SVM procedure alongside MapReduce worldview to accomplish a higher precision in recognition of the spam email. In [6] built up a version as a solution for spotting phishing websites by means of utilising the URL popularity approach utilising the Random forest algorithm. There are three widespread stages like Parsing, Heuristic type of statistics, overall performance evaluation in this model and each stage utilizes an exchange method the calculation for dealing with information to give higher results. In [7] proposed another Software-Defined Networking based region approach that utilizes the ascribes of ransomware similitudes. Taking into account the view of the association, the likenesses between the ransomware families, explicitly CryptoWall and Locky, they accepted that an appraisal of the HTTP message plans and their substance sizes is satisfactory to see such dangers. They showed the chance of our procedure by arranging and surveying a proof-of-thought SDN-based disclosure structure. The test outcomes confirm that the proposed approach is attainable and capable. In [8] studied different strategies for digital assaults, endeavours to relieve them, their solidarity, and shortcoming. They additionally examined the conventional engineering of URL boycotting that is being utilized by the different malware identification frameworks. In [9] proposed a learning-based accumulation examination component to choose page design likeness, which is utilized to identify phishing pages. They model their answer and assess four mainstream AI classifiers on their precision and the components influencing their outcomes. In [10] proposed the strategy to apply various philosophies for seeing phishing messages utilizing suggested likewise as new highlights. A few novel information consolidates that can help with finding phishing assaults with especially limited a-earlier information about the foe or the strategy used to dispatch a phishing assault. Their way of thinking is to bundle phishing messages by joining key fundamental highlights in phishing messages. In[11] Phishing has gotten a standout amongst the alternative three most cutting-edge types of law-breaking in keeping with boost reviews, and the two frequencies of activities and client defencelessness have elevated as of overdue, definitely consolidating the peril of financial underhandedness.In[12] The essential key element is to permit the client to ask whether visited sites are unique or phony. In[13] A multiple classification algorithms are utilized which incorporates SVM, AdaBoost, and Naive Bayes. These calculations are separated into three levels utilizing 21 fixed at this point various highlights. At that point a two-venture strategy happens with the assistance of another classification calculation yet the issue here is the time burned-through and the intricacy in question, overheads included, and the presentation issues and consequently this was anything but an ideal strategy.

II. Proposed Solution

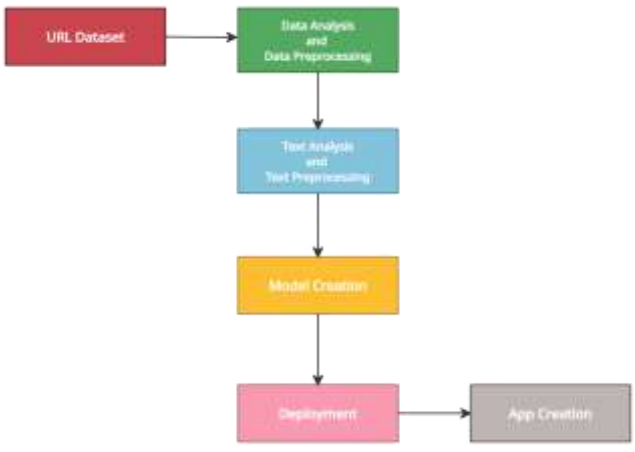

Fig. 1 Data Flow

Fig. 1, shows how the raw data gets transformed into valuable data by performing EDA and data pre-processing. The dependent attribute and independent attributes are then made to

fit into the model. Then the best model is deployed and is used for the app creation. In the model creation, we used two algorithms 1. Multinomial Naïve Bayes[14] 2. Logistic Regression.

A. Multinomial Naive Bayes: Multinomial Naive Bayes algorithm best works on text data. It counts the probability of each word in a text document. It counts for multiple repetitions of words. It uses common words differently as compared to usual ones. It is faster than the Naive Bayes algorithm.

The Naive Bayes algorithm comes from the Bayes theorem and it can be formulated in Eqn-1.

$$P(A|B) = \frac{P(B/A) \times P(A)}{P(B)} \qquad \text{Eqn-1}$$

We can write this equation in terms of X(input variable) and y(output variable) as shown in Eqn-2.

$$P(y \ / \ X) = \frac{P(X|y) \times P(y)}{P(X)} \qquad \text{Eqn-2}$$

For multiple inputs and because of naive assumptions for the independent variable. We can write P(X|y) as shown in Eqn-3.

$$P(y) = P(y) * P(y) * \ldots * P(x_n|y) \qquad \text{Eqn-3}$$

Additionally, since we are settling for y, P(X) is consistent which implies that we can eliminate it from the condition and present a proportionality. The objective of Naive Bayes is to pick the class y with the most extreme likelihood. Argmax is essentially an activity that discovers the contention that gives the most extreme worth from an objective capacity. For this situation, we need to track down the most extreme esteem.

$$y = argmax_y \ [P(y) * \prod_{i=1}^{n} P(x_i|y) \qquad \text{Eqn-4}$$

With the help of the Eqn-4, we can find the probability of each word and detect the phishing nature of URLs.

B. Logistic Regression: Logistic Regression is a generalized linear model. Unlike Linear Regression, which is used to predict the values according to numeric data, Logistic Regression is used for problems. the **classification.** Some of the examples are Email spam classifiers, iris species classification, credit card fraud detection, and Fake News classification.

Logistic Regression predicts the probability value of the dependent feature which ranges between [0,1]. If it is a binary classification problem and if the value is greater than or equal to 0.5 then it is classified as true else, it is classified as false. However, we know that in Linear Regression, the range is from negative infinity to positive infinity, but here we have a value between [0,1]. To solve this problem, we have a function known as the **Sigmoid Function.** The Sigmoid function gives a squiggly line.

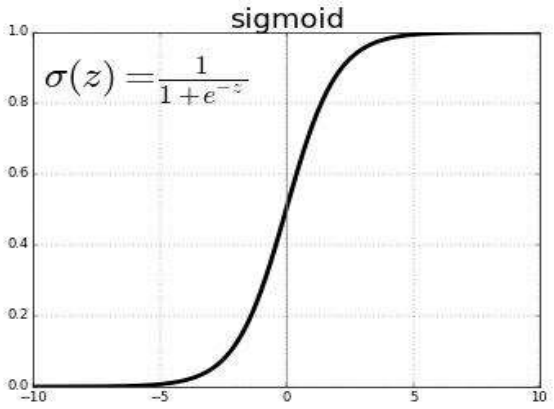

Fig. 2 Graphical representation of a sigmoid function

B.1 Maximum Likelihood estimation: In Linear Regression, we use Least squares which is the sum of the square of the error to find the best fit line. In figure 2, as most of the points tend to negative infinity and positive infinity, so the error value will also tend to infinity. We cannot use Least square, instead, we use **Maximum Likelihood estimation** using Eqn-5**.**

$$p = \frac{e^{log\ (odds)}}{1+e^{log\ (odds)}} \qquad \text{Eqn-5}$$

We calculate the maximum likelihood of every log(odds) point and multiply all likelihood to get the likelihood of a complete dataset. The squiggly curve which gets the maximum likelihood value of the complete dataset is considered as the best squiggly curve.

B.2 Cost function in Logistic Regression: The expense work in Logistic Regression addresses enhancement objectives, i.e. the production of an expense capacity and minimization of cost work utilizing Gradient Descent with the end goal that worldwide minima can be accomplished.

$$J(\theta) = \frac{1}{2m}\sum \quad (h_\theta(x^{(i)}) - y^{(i)})^2 \qquad \text{Eqn-6}$$

So, if we try to use the cost function of linear regression used in Eqn-6 for the hypothesis function(sigmoid equation) of logistic regression then it is observed that it gives a nonconvex function[Figure-3]. In the non-convex function, we get the local minimum in addition to the global minimum, and finding the global minimum will be a difficult task.

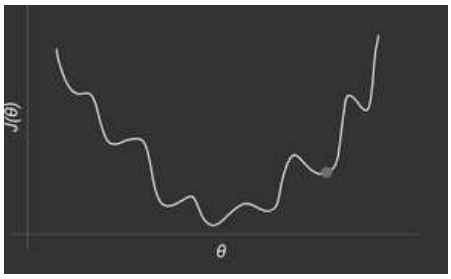

Fig. 3 Cost function in logistic regression

For Logistic Regression, the Cost function will be defined in Eqn-7: -

$$J(\theta) = -\frac{1}{m}\sum \quad [y^{(i)} log(h_\theta(x^{(i)}) + (1 + y^{(i)}) log(1 - h_\theta(x^{(i)})] \qquad \text{Eqn-7}$$

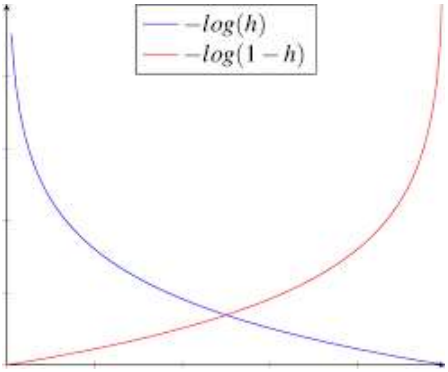

Fig. 4 Graphical representation of cost function

C. Model Creation: The dataset used is taken from Kaggle which consists of approximately 5 Lakh unique rows and two columns, the URLs and the Label. The first step is always to find meaningful insights from the data like the dataset does not contain any NaN values and the dataset was not imbalanced. After data analysis, data preprocessing is done where categorical labels are converted into the numerical format by using LabelEncoder. After the data preprocessing, now the text data is analyzed by using a word cloud. The URLs are tokenized using the tokenizer and the tokenized words are stemmed using snowball stemmer. Stemmer processes fast as compared to lemmatization. Some words are not necessary for model creation and need to be removed, known as stop words like “the”, “a” “are”, “in”. Then the stemmed words are grouped to make sentences. The stemmed sentences need to be converted into the numeric form using TF-IDF (Term Frequency and Inverse Document Frequency). It will not only give the numerical value based on the number of occurrences but also provide importance to that word. It creates a matrix that contains the fraction value. Now we create a model, since it’s a classification problem, logistic regression is used for model creation. Multinomial Naïve Bayes gives better accuracy with text data and it works with the data which contains the frequency of each word in the Text Document. We found that logistic regression provides 96.40 % accuracy whereas MultinomialNB gave 95.76% accuracy so we use logistic regression to dump our model. You can see the comparison of both models in Fig. 5.

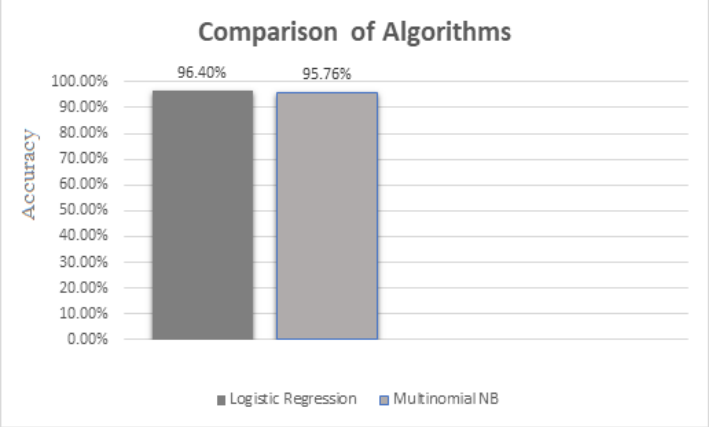

Fig. 5 Comparison of Algorithms

Next, we create a pipeline and dump our model using pickle and get a .pkl file as an output. Now we deploy our model in Heroku and use the flask to create an app while checking the resultant of the site shown in Figure-6.

Fig. 6 Result on the phishing site.

3. Results and Discussion

The dataset is taken from Kaggle which consists of 5 Lakh entries that contain two columns (URLs and Label). We discovered that it contains some null values which need to be removed. The dataset was imbalanced i.e., the proportion of phishing site and the legit site was not balanced. We solved this problem by using K-Fold cross-validation. In-text pre-processing, we found it contains a lot of stop words that need to be removed as these words do not contribute during model creation. After tokenization and stemming, the bag of words needs to be converted into numerical form. We used TF-IDF (Term Frequency-Inverse document Frequency) which performed better than the count vectorizer. We used Logistic Regression which is used for classification problems and Multinomial Naive Bayes which performs better with text data and found that Logistic Regression performed better than MultinomialNB with an accuracy of 96.40% and 95.74%. So the pipeline of Logistic Regression is created and then dumped using pickle. an open undertaking for further studies and development. This file is then deployed on Heroku. This examination gives an unrivalled perception of phishing locales. We used distinctive data pre-processing and text pre-processing techniques to improve the precision.

4. Conclusion

Phishing is a shocking danger in the web security area. In this assault, the client inputs his/her data to a fake site that resembles a real one. We have introduced an overview of phishing recognition approaches by utilizing NLP and Machine Learning methods. This study gives a superior comprehension of phishing sites. We utilized different information pre-processing and text pre-processing strategies to improve precision. We got 96.40 % exactness for Logistic relapse and 95.76 % precision for MultinomialNB. We made a pipeline utilizing calculated relapse for Logistic regression and unloaded

the model utilizing the Pickle. Finally, we conveyed the model in Heroku and made an API utilizing the flask.

References:

[1] Mohammed HazimAlkawaz, Stephanie Joanne Steven, Asif Iqbal Hajamydeen," Detecting Phishing Website Using Machine Learning", 2020 16th IEEE International Colloquium on Signal Processing & its Applications (CSPA 2020), 2 8-29 Feb. 2020, Langkawi, Malaysia

[2] Bryan Espinoza∗, Jessica Simba, Walter Fuertes, Eduardo Benavides, Roberto Andrade, and Theofilos Toulkeridis, "Phishing Attack Detection: A arrangement dependent on common Machine picking up demonstrating cycle", 2019, International Conference on Computational Science and Computational Intelligence (CSCI)

[3] Mahajan Mayuri Vilas, Kakade Prachi Ghansham, Sawant PurvaJaypralash, Pawar Shila," Detection of Phishing Website Using Machine Learning Approach", 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)

[4] Prajakta Patil, Rashmi Rane, Madhuri Bhalekar," Detecting Spam and Phishing Mails Using SVM and Obfuscation URL Detection Algorithm",2017 International Conference on Inventive Systems and Control (ICISC).

[5] Arun Kulkarni, Leonard L. Brown, "Phishing Websites Detection using Machine Learning", International Journal of Advanced Computer Science and Applications 10(7), 2019.

[6] Purvi Pujara, M. B. Chaudhari, "Phishing website detection using machine learning: A Review", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN: 2456-3307, Volume 3 Issue 7, pp. 395-399, September-October 2018.

[7] Huaping Yuan, Xu Chen, Yukun Li, Zhenguo Yang, and Wenyin Liu, "Detecting Phishing Websites and Targets Based On URLs and Webpage Links", 2018 24th International Conference on Pattern Recognition(ICPR) Beijing, China, August 20-24, 2018.

[8] Martyn Weedon, Dimitris Tsaptsinos and James Denholm-Price, "Random Forest Explorations for URL Classification", 2017

[9] Rohith S, Sujatha B K, " URL Detection using Combined key sequence of Logistic map and Lozi map," Proc. of International Conference on Communications and Signal Processing (ICCSP), Tamilnadu, 2015, pp. 1053-1058.

[10] Al-Najjar, Hazem Mohammad, Asem Mohammad AL-Najjar, and K. S. A. Arar. " Phishing Websites algorithm based on logistic map and pixel mapping table" Proc. ofInternational Arab Conference on Information Technology, (ACIT 2011), pp. 56-60. 2011.

[11] Liu, Jing-mei Qu, Qiang, "Phishing Website Detection using Machine Learning", Proc. of the Third International Symposium on Information Processing, October 2010, pp 67 - 69.

[12] Hong, Lianxi Li, Chuanmu. "Phishing Websites Detection using Machine Learning", Proc. of IEEE 2nd International Conference on Anti-counterfeiting, Security, and Identification, 2008, pp 1-5.

[13] Safi, Haifaa W, Maghari, Ashraf Y. A, "Detecting Spam and Phishing Mails Using SVM and Obfuscation URL Detection Algorithm,2" Proceedings of IEEE International Conference on Promising Electronic Technologies (ICPET), 2017, pp. 66-70.

[14] Eitel J. M. Lauría, Alan D. March. "Combining Bayesian Text Classification and Shrinkage to Automate Healthcare Coding", Journal of Data and Information Quality, 2011.

[15] Ranjith, R., Supreeth, S., Ramya, R., Ganesh Prasad, M., Chaitra Lakshmi, L., "Password processing scheme using enhanced visual cryptography and OCR in hybrid cloud environment", International Journal of Engineering and Advanced Technology, 2019, 8(5 Special Issue), pp. 150–154.

[16] Supreeth S, P Sarika, Shweta Kumari, Seema M, Sona Singh "Framework for data security from SQL Injection in cloud computing", International Journal of Advanced Research in Computer Science, Vol:9, Issue:3, PP: 257-262, 2018.