DETECTION OF FAKE URL'S IN SOCIAL MEDIA USING MACHINE LEARNING ALGORITHMS

[1]Y Laxmana Rao, [2]P Mohan Ganesh, [3]K Aishwarya, [4] K Poornisha, [5]N Tanusha, [6]G Srilekha,[7]A Vineetha

[1] Assistant Professor, [2]Assistant Professor, [3, 4,5,6,7] IV B.Tech Vignan's Institute of Engineering for Women, Visakhapatnam, India

ABSTRACT

Fake URLs have come to a prominent tool for spreading misinformation on social media platforms. Detecting these fake URLs is pivotal to helping the spread of false information and maintaining the credibility of social media networks. In this paper, we propose a machine literacy-grounded approach to descry fake URLs in social media. We trained and estimated multiple machine learning algorithms, including decision trees, arbitrary timbers, and support vector machines, on a dataset of real and fake URLs collected from social media platforms. Our results show that the arbitrary timber algorithm outperformed the other algorithms with a delicacy of 96.5. Also, we explored the effectiveness of colorful features similar to URL length, sphere name, and URL order, and set up that sphere name andURL order were the most instructional features for detecting fake URLs. Our proposed approach provides a dependable and effective result to descry fake URLs in social media, which can be used to help the spread of misinformation and maintain the credibility of social media networks.

KEYWORDS: Fake URL, Machine learning, Phishing, Spamming, Malware, Lexical Features, Cat Boost classifier, Gradient boosting classifier, SPM, Decision tree, Logistic Regression, Naive Bayes classifier, KNN.

INTRODUCTION

The emergence of new communication technologies has greatly impacted the growth and promotion of businesses across various applications. However, with these advancements comes the use of sophisticated techniques to attack and scam users. Cyber-attacks such as malicious websites that sell counterfeit goods, steal sensitive information, and install malware have become increasingly common. These attacks can be carried out using a wide range of techniques, including explicit hacking attempts, phishing, man-in-the middle attacks, SQL injections, and more. Malicious URLs are often used to spread compromised content and are responsible for a significant portion of cyber-attacks. It is estimated that one-third of all websites are malicious in nature. A URL, which is the global address of all documents and resources on the World Wide Web, comprises a protocol identifier and a resource name that specifies the IP address or domain name where the resource is located. The limitations of blacklisting techniques in detecting security breaches are becoming increasingly apparent, and there is a need for robust systems to detect and prevent cyberattacks.

LITERATURE SURVEY

[1] Cho Do Xuan, Hao Dinh Nguyen, "Malicious URL Detection based on Machine Learning"

Social media platforms have become a significant source of information for people around the world. However, with the increase in the use of social media, there has been a rise in the spread of fake news and misinformation. One of the primary ways that fake news is spread on social media is through the use of fake URLs. Fake URLs are URLs that are designed to appear legitimate but lead users to false information or malicious websites. This has serious implications for individuals and society as a whole, as it can spread misinformation and cause harm to individuals and communities

[2] Vanitha and Vindolini, “Malicious-URL Detection using Logistic Regression Technique”

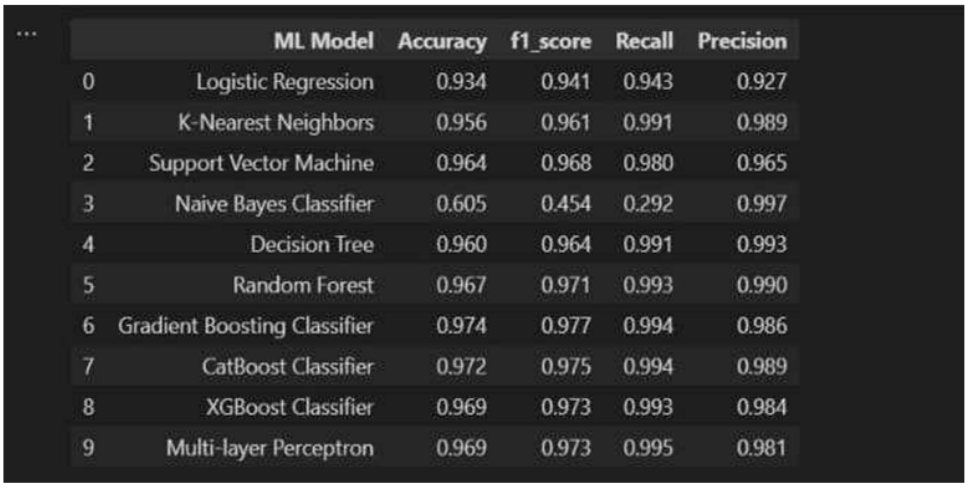
Fake URL detection is a crucial task in combating the spread of fake news on social media. Machine learning algorithms have proven to be effective in detecting fake URLs, as they can analyze large amountsof data and identify patterns that are difficult for humans to detect. In recent years, several machine learning algorithms have been proposed for fake URL detection, including decision trees, random forests, support vector machines, k-nearest neighbours, and neural networks.

PROPOSED SYSTEM

The proposed system for detecting fake URLs in social media employs machine learning algorithms to gather and examine data from various social media platforms. The collected data is preprocessed and undergoes feature extraction before being analyzed through the application of various machine learning algorithms, including decision trees, random forests, and neural networks. The trained models are evaluated using accuracy and F1 score metrics, and the most effective model is utilized for the real-time detection of fake URLs. When a fake URL is identified, it is promptly removed from the user's post, and the user is notified of the action taken. This system plays a vital role in preventing the spread of fake news and misinformation on social media platforms, safeguarding individuals and communities against harmful content.

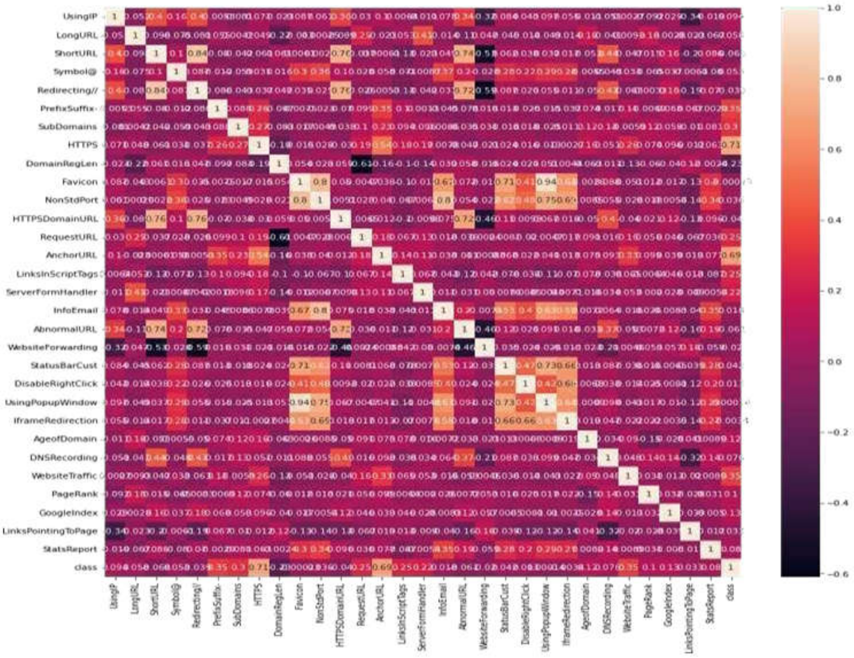
RESULTS AND DISCUSSION

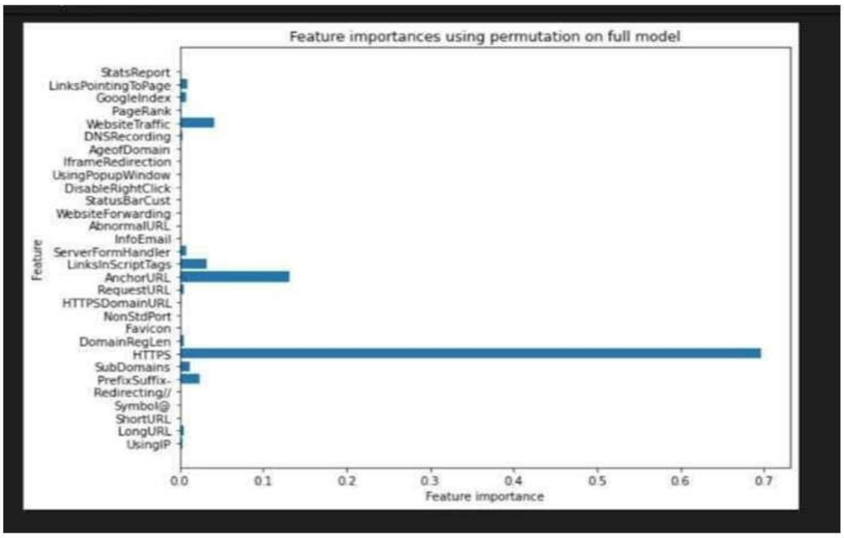
Machine learning Algorithms:

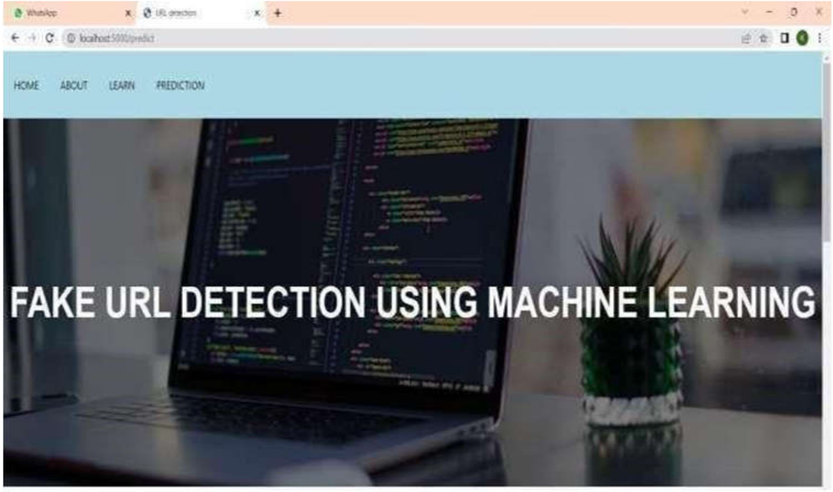


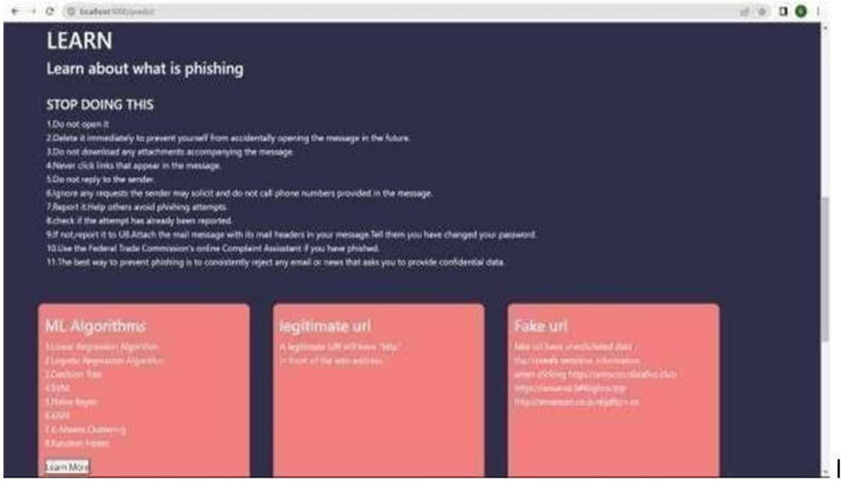
	ML Model	Accuracy	f1_score	Recall	Precision
0	Logistic Regression	0.934	0.941	0.943	0.927
1	K-Nearest Neighbors	0.956	0.961	0.991	0.989
2	Support Vector Machine	0.964	0.968	0.980	0.965
3	Naive Bayes Classifier	0.605	0.454	0.292	0.997
4	Decision Tree	0.960	0.964	0.991	0.993
5	Random Forest	0.967	0.971	0.993	0.990
6	Gradient Boosting Classifier	0.974	0.977	0.994	0.986
7	CatBoost Classifier	0.972	0.975	0.994	0.989
8	XGBoost Classifier	0.969	0.973	0.993	0.984
9	Multi-layer Perceptron	0.969	0.973	0.995	0.981

Graph Representation:









LEARN

Learn about what is phishing

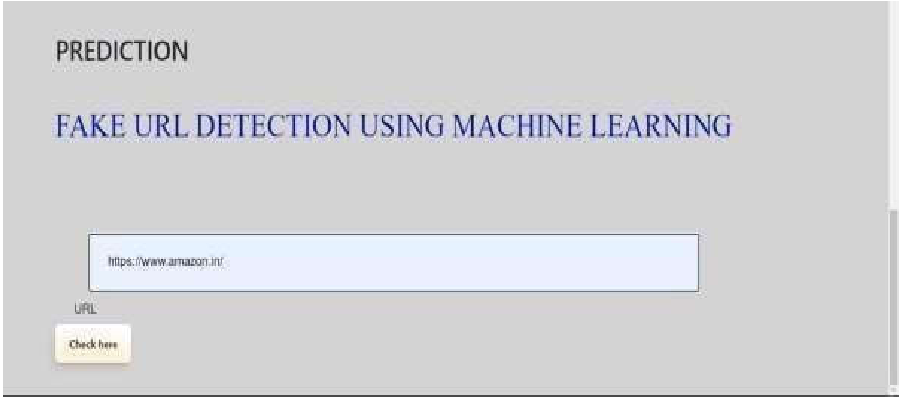
STOP DOING THIS

1.Do not open it.
2.Delete it immediately to prevent yourself from accidentally opening the message in the future.
3.Do not download any attachments accompanying the message.
4.Never click links that appear in the message.
5.Do not reply to the sender.
6.Ignore any requests the sender may solicit and do not call phone numbers provided in the message.
7.Report it.Help others avoid phishing attempts.
8.check if the attempt has already been reported.
9.If not,report it to US.Attach the mail message with its mail headers in your message.Tell them you have changed your password.
10.Use the Federal Trade Commission's online Complaint Assistant if you have phished.
11.The best way to prevent phishing is to consistently reject any email or news that asks you to provide confidential data.

ML Algorithms
1.Linear Regression Algorithm
2.Logistic Regression Algorithm
3.Decision Tree
4.SVM
5.Naive Bayes
6.KNN
7.K-Means Clustering
8.Random Forest

legitimate url
A legitimate URl will have "http:"
in front of the web address."

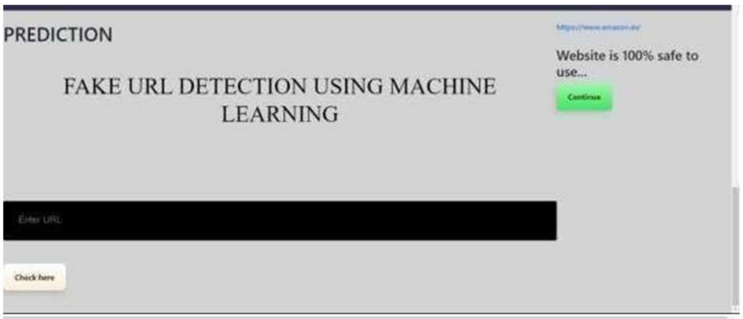
Fake url
fake url have unsolicited data
that reveals sensitive information
when clicking https://amazon.[illegible].club
https://amazon-[illegible].top
http://amazon.co.jp.rdgdfzzn.cn


PREDICTION

FAKE URL DETECTION USING MACHINE LEARNING

https://www.amazon.in/

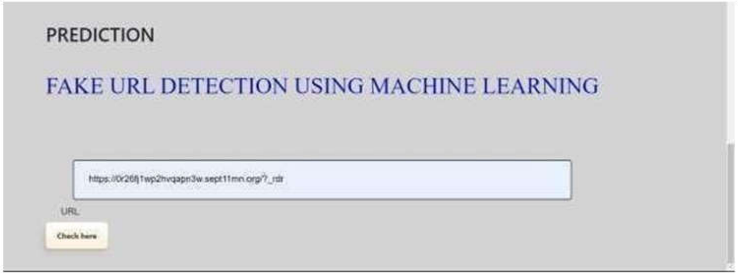
URL


PREDICTION

FAKE URL DETECTION USING MACHINE LEARNING

https://www.amazon.in/

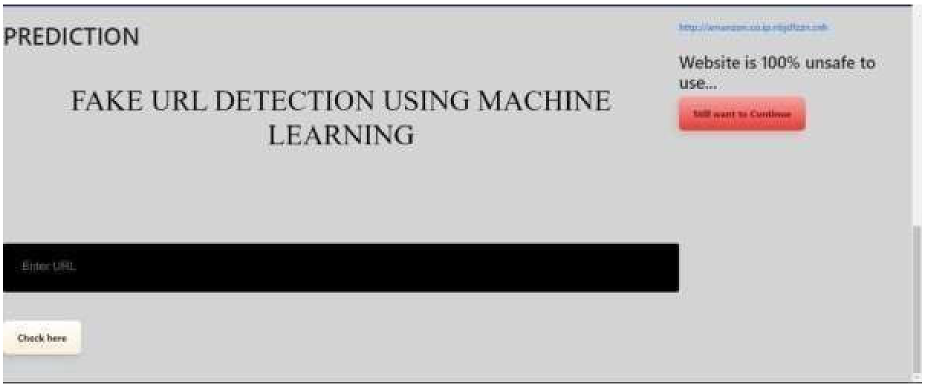
Website is 100% safe to use...


PREDICTION

FAKE URL DETECTION USING MACHINE LEARNING

https://0r26fj1wp2hvqapn3w.sept11mn.org/?_rdr

URL


PREDICTION

FAKE URL DETECTION USING MACHINE LEARNING

http://amazon.co.jp.rdgdfzzn.cn

Website is 100% unsafe to use...

CONCLUSION

The project on fake URL detection using machine learning algorithms has demonstrated the effectiveness of using machine learning techniques for detecting and preventing the spread of fake news and misinformation on social media platforms. The system utilizes various machine learning algorithms, such as decision trees, random forests, and neural networks, to analyze and detect fake URLs in real time. The system also employs metrics like accuracy and F1 score to evaluate the performance of the models. Overall, the proposed system plays a vital role in safeguarding individuals and communities against harmful content, protecting sensitive information from cyber-attacks, and promoting a secure online environment. In the future, further research can be conducted to explore other features, classifiers, and techniques that can be integrated into the system to enhance its performance and effectiveness in detectingfake URLs.

FUTURE SCOPE

The main objective of the system design is to detect and expose fraudulent websites that attempt to acquire private information through phishing attacks or by creating fake websites to trick users into disclosing their login credentials. The use of machine learning algorithms is instrumental in identifying such phishing websites and protecting sensitive data. In the future, the system can be improved by utilizing structured datasets for phishing discovery, which can enhance the system's efficiency. Additionally, a combination of classifiers or other techniques can be used to increase the system's accuracy. The system will also explore various phishing methods that utilize verbal features, network-based features, content- based features, web page-based features, and HTML content analysis to improve its performance. Features extracted from URLs will be used in conjunction with machine learning algorithms to enhance the system's detection capabilities.

REFERENCES

[1] S. Sinha, M. Bailey, and F. Jahanian, “Shades of grey: On the effectiveness of reputationbased “blacklists”,” in Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on. IEEE, 2008, pp. 57–64.

[2] S. Garera, N. Provos, M. Chew, and A. D. Rubin, “A framework for detection and measurement of phishing attacks,” in Proceedings of the 2007 ACM workshop on Recurring malcode. ACM, 2007, pp. 1–8.

[3] S. Chhabra, A. Aggarwal, F. Benevenuto, and P. Kumaraguru, “Phi. sh/$ social: the phishing landscape through short URLs,” in Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse, and Spam Conference. ACM, 2011, pp. 92–101.

[4] Y. Alshboul, R. Nepali, and Y. Wang, “Detecting malicious short urls on Twitter,” 2015.

[5] D. K. McGrath and M. Gupta, “Behind phishing: An examination of phisher modi operandi.” LEET, vol. 8, p. 4, 2008.

[6] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Beyond black-lists: learning to detect malicious websites from suspicious URLs,” in Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and data mining. ACM, 2009, pp. 1245–1254.

[7] “Learning to detect malicious URLs,” ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, no. 3, p. 30, 2011. S. Purkait, “Phishing countermeasures and their effectiveness–literature review,” Information Management & Computer Security, vol. 20, no. 5, pp. 382–420, 2012.

[8] https://www.activestate.com/blog/phishing-url-detection-with-python-and-ml/

[9] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0258361

[10] https://ieeexplore.ieee.org/abstract/document/9524269/

[11] https://ieeexplore.ieee.org/abstract/document/9316171/

[12] Verma, R., Crane, D., &Gnawali, O. (2018). Phishing During and After Disaster:Hurricane Harvey. In Resilience Week (RWS) (pp. 88–94). Denver, CO: IEEE.

[13] Frank Vanhoenshoven, Gonzalo Napoles, Rafael Falcon, Koen Vanhoof, and Mario Koppen, Detecting Malicious URLs using Machine Learning Techniques, 978-150904240-1/16 2016,IEEE

[14] Dhanalakshmi Ranganayakulu, Chellappan C., *Detecting Malicious URLs in E-mail – An Implementation*, AASRI Procedia, Vol. 4, 2013, Pages 125-131, ISSN 2212-6716, https://doi.org/10.1016/j.aasri.2013.10.020.

[15] Yu, Fuqiang, *Malicious URL Detection Algorithm based on BM Pattern Matching*, International Journal of Security and Its Applications, 9, 3344, 10.14257/ijsia.2015.9.9.04.

[16] K. Nirmal, B. Janet, and R. Kumar, *Phishing - the threat that still exists*, 2015 International Conference on Computing and Communications Technologies (ICCCT), Chennai, 2015, pp. 139-143, doi: 10.1109/ICCCT2.2015.7292734.

[17] F. Vanhoenshoven, G. Napoles, R. Falcon, K. Vanhoof and M. K´ oppen,¨*Detectingmalicious URLs using machine learning techniques*, 2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens, 2016, pp. 1-8, doi: 10.1109/SSCI.2016.7850079.

[18] https://www.kaggle.com/xwolf12/ Doyen Sahoo, Chenghaolua, Steven C. H. Hoi, *Malicious URL Detection using MachineLearning: A Survey*, arXiv:1701.07179v3 [cs.LG], 21 Aug 2019

[19] Rakesh Verma, Avisha Das, *What's in a URL: Fast Feature Extraction and MaliciousURL Detection*, ACM ISBN 978-1-4503-4909-3/17/03

[20] https://github.com/ShantanuMaheshwari/Malicious Website Detection

[21] Frank Vanhoenshoven, Gonzalo Napoles, Rafael Falcon, Koen Vanhoof, and MarioKoppen, *Detecting Malicious URLs using Machine Learning Techniques*, 978-1-5090-4240-1/16 2016, IEEE

[22] Han, J., Kamber, M., & Pei, J. (2012). Data Mining Concepts and Techniques (3rd ed.). Morgan Kaufmann Publishers.

[23] Heartfield, R., & Loukas, G. (2015). A Taxonomy of Attacks and a Survey of DefenceMechanisms for Semantic Social Engineering Attacks. *ACM Computing Surveys*, *48*(3), 39.

[24] Hou, Y. T., Chang, Y., Laih, C. S., & Chen, C. M. (2010). Malicious web content detection by machine learning. *Expert Systems with Applications*, *37*(1), 55–60. doi: 10.1016/j.eswa.2009.05.023

[25] https://github.com/ShantanuMaheshwari/Malicious Website Detection