Malicious URL Detection Using Machine Learning

[1]Vivek R. Tiwari, [2]Shivam H. Mishra, [3]Sourabh S. Tiwari,[4]Poonam Talele
[1]Computer Science,
[1]Shivajirao s. jondhale college of engineering, Mumbai, India

Abstract : In Cybersecurity there are various kinds of threats, but Malicious URL is one of the biggest threat. Malicious URL contains many harmful data like Spam ,Virus, Trojan etc. due to which people faces loss of Money and confidentiality. It is quite difficult to detect and overcome from these problems within the time. Usually to detect this people use JavaScript code, Blacklist SEO techniques. However, blacklists are not vast enough to detect latest malicious URL. Whereas the Scope of JavaScript code is limited. Now a days Machine Learning is considered to be one of the best method to improve the generality of malicious URL detectors. Therefore, in our project we are using machine learning algorithm like tokenization, vectorization and logistic regression for better result.

I Introduction : The arrival of new communication technologies has high impact in the success of the businesses across a lot of applications including e-commerce, online-banking and social networking. In fact, at present it is quite necessary to have the online presence to successfully run a Business. Unfortunately, the technological advancements come together with new sophisticated techniques to scam and attack naive users. These kinds of attacks contains duplicate websites that sell counterfeit goods, fraud of money by tricking users into obtaining confidential information which ultimately leads to loss of money or identity of users, or even installing malicious contents in the user's system. There are various methods to perform such attacks, like drive-by exploits, social engineering, man-in-the middle, phishing, explicit hacking attempts, creating loophole, SQL injections, loss or theft of devices, DDOS(distributed denial of service), and a lot of others. Considering the various types of attacks, probably new types of attacks, and the endless contexts in which such kinds of attacks may appear, it is difficult to design durable systems to detect cyber-security violations. The limitations of security management technologies that are being used at present are becoming more and more grave given the tremendous growth of latest security threats, rapidly changing new IT technologies, and considerable shortage of cyber security officials. Distributing compromised URLs are used to realize most of these attacking techniques.

II Literature Survey : Narendra. M. Shekokar et al. have given one of the Approach of detection and avoiding the phishing attacks, which is ultimately focused on proceeds to the visual identical-based detection, Link Guard Algorithm is used for analyzing the two URLs and finally based on the result given by the algorithm the procedure processes to the next stage. High false positive is the drawback of this paper.

M. Rajesh et al. have explained URL Attacks , which basically focusses on searching Malicious URLs. Conditional redirection technique was used by which the URLs get categorized and the target page as well that the user needs are satisfied. For differentiating the URLs the learning techniques are also introduced. Dynamic redirections is the main drawback of this paper.

Bargal Varsharani et al. have explained Detecting Suspicious URLs using Bayesian Classification. Discovering lexical and host-based properties of malicious web sites is the main application. So the approach it uses to prevent the system is dependent on URL classification, using statistical techniques to find lexical and host-based competency of spam URLs. Even though, detecting wide range of malicious URLs.

Roshani et al. have introduced ML system to detect spam URLs and malicious contents and to find out if a particular tweet is malicious or not in the Social Media site like Facebook etc. We classified the input tweet by collecting dataset and training the classifier. The Naive Bayes algorithm is a supervised learning model with companion learning algorithms that are being used to analyze data used for regression and classification analysis. The sensitivity of each tweet was calculated after classification.

Chia et al. have introduced the Feature identification for finding spam URLs using Bayesian classification in social media sites. Here, a feature set was shown that couple the features of traditional social networking and heuristics. Furthermore, based on Bayesian classification a spam URL detection system to use in social sites was proposed. Results reflects that the proposed approach achieves a high detection rate Experimentally.

III Problem definition : In this project we are going to deal with the problem of avoiding malicious website. Here we are going to differentiate between malicious and non-malicious website by comparing it with the parameters of a non-malicious/malicious URL data sets. We are going to make sure that the dignity of the users is maintained. We will make sure that we provide more accuracy, usability, security as compared to the previous existing systems. In this project for attaining accuracy, usability and security, we are going to use the following algorithm's stated as Tokenization, Vectorization and Logistic Regression.

IV Existing System : As shown in the Figure 4.1 model briefly explains the path that they have followed in their project. It initially begins by collecting model data where the data are taken from various social networking sites like Twitter, Facebook, and followed by feature extraction and labelling of URLs, followed by classification and the end result of the classification algorithm. The below section gives summarized explanation of each and every module in detail.

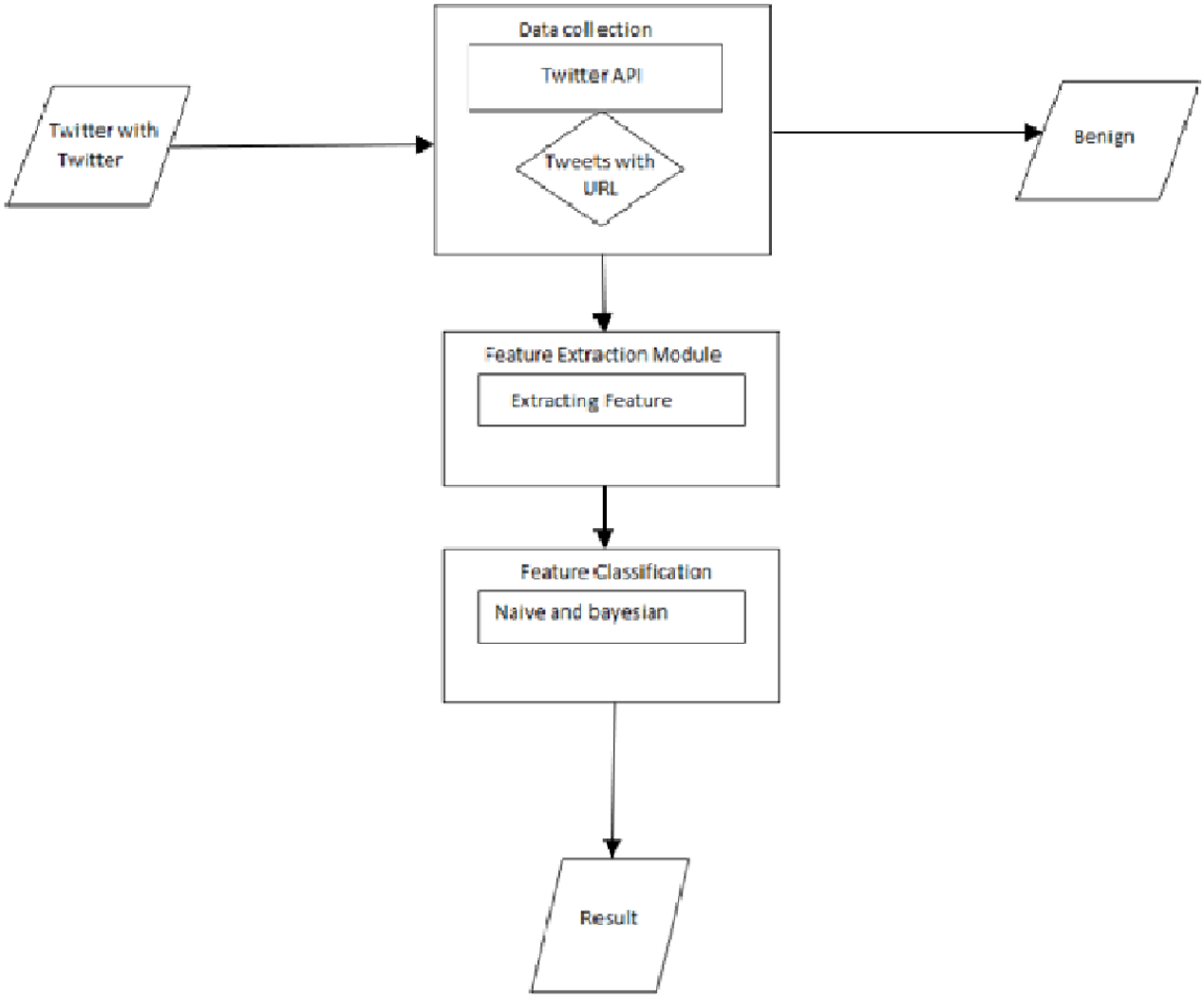
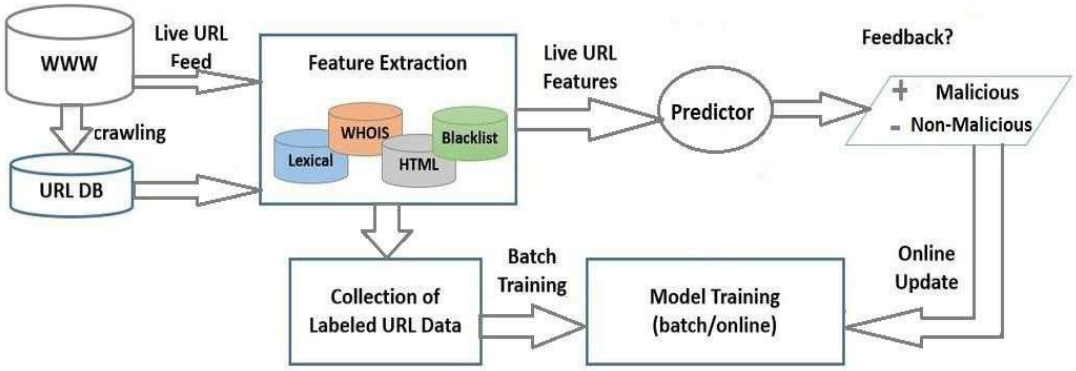


Fig 4.1. Existing System

Drawbacks

1. In this system it only focuses on social networking system i.e. it has limited scope.
2. Naive Bayesian algorithm is little bit complex due to which it's difficult to do.

V Proposed System : The proposed system basically consists of client side and server side. On client side, there can be any computer. On client-side users will surf the web pages using any browser. On server-side server will perform following tasks. As shown in fig 5.1



5.1 Architecture Diagram

ML Approach

Machine Learning approach basically consists of two things:

Datasets

Machine Learning model

The Model is to be sketched in such a way that the meaningfully information is perceived in the simplest way and then from that it is to be trained upon and trying to develop a defined behavior from the collected data-sets. Data-sets are basically the backbone of any model and therefore it should be sufficient and good enough data for safe as well as unsafe URLs available in the data-sets for the model to be trained upon.

Modeling the system

The ML approach for detecting the malicious URLs can be first involve in improvement of our information within the data-sets. We will define some vectorizer to clear the data-sets and use pandas. More on this later. We are using Logistic Regression Algorithm to train our model. Vectorization can be understood by following fig 5.2

Id	Message
123	"be happy"
321	"to be or not to be"

Id	"be"	"happy"	"be happy"	"to"	"or"	"not"	"to be"	"be or"	"or not"	"not to"
123	1	1	1	0	0	0	0	0	0	0
321	1	0	0	1	1	1	1	1	1	1

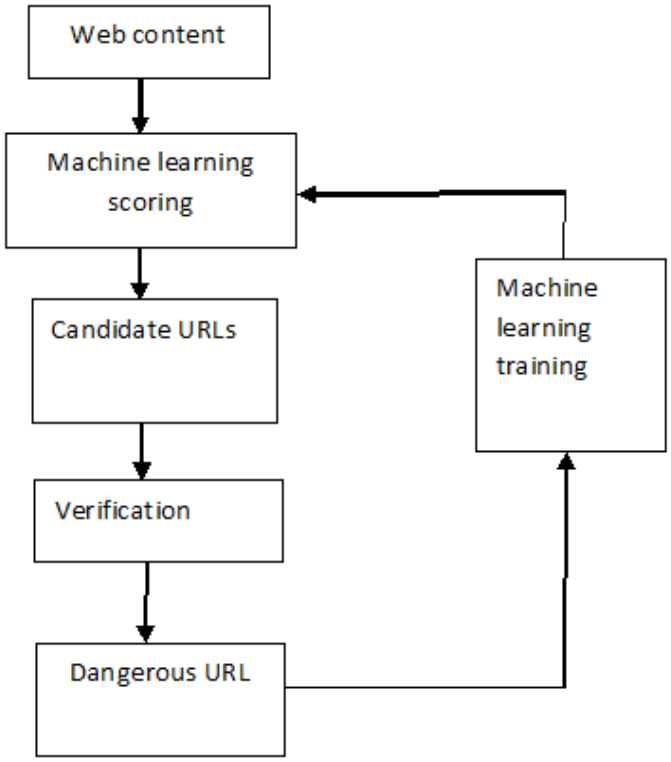
5.2 vectorization example

Logistic Regression : Logistic Regression algorithm is basically binary classification method and hence it is also called as Binary Regression. This model is dichotomous and will give result as : no/yes. This is done by using Utmost probability estimation. Simply the highest value will be used to optimize the colligation values for training the data. The key difference from simple regression is that the output price value is in the form of binary (0 or 1) rather than numeric.

Preparing Data : As we know that the URLs are quite different from traditional text documents, we need to define some sanitization technique to get out the required data from raw URL data. We will implement defined sanitization function in python in order to filter the URLs.

Training the model : Now the interesting part is that as we have stated above we will use logistic regression to train our model but before that we need to pass the data to custom defined vectorizer.

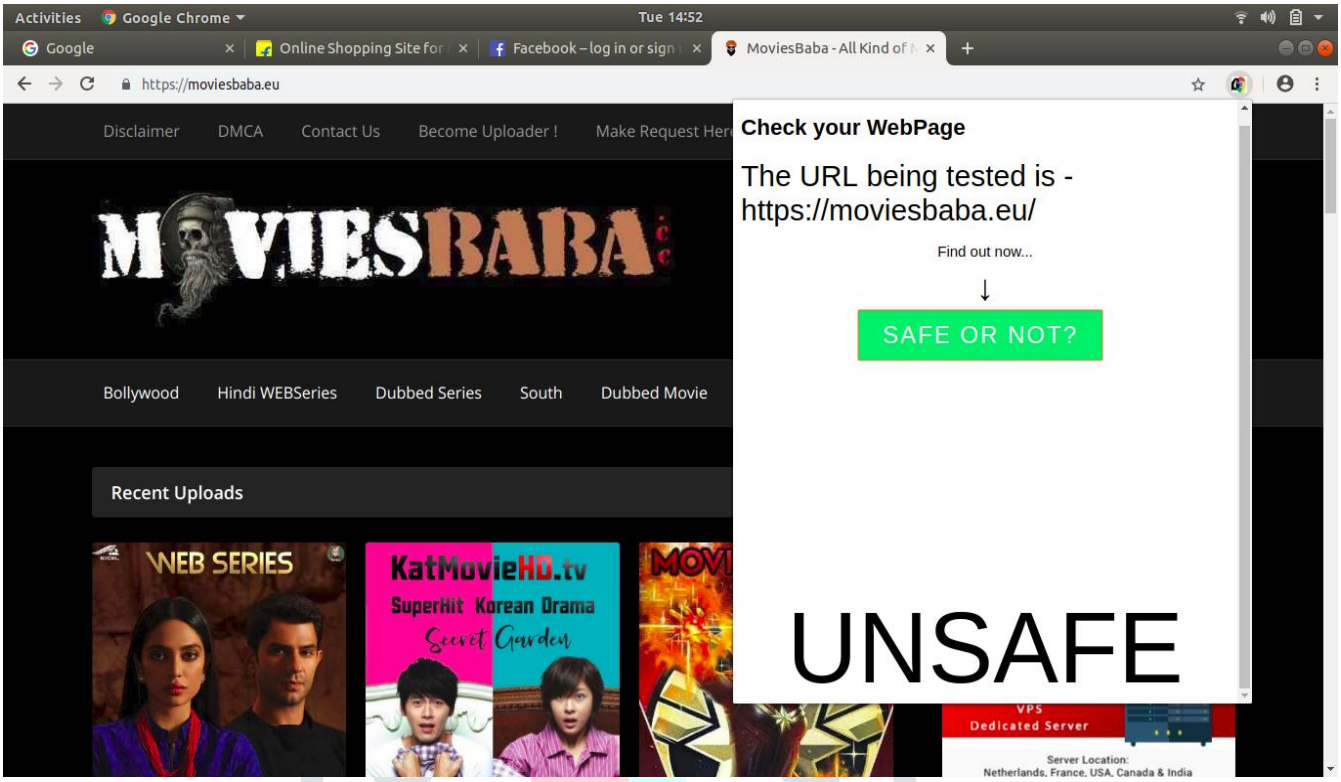
VI System Design

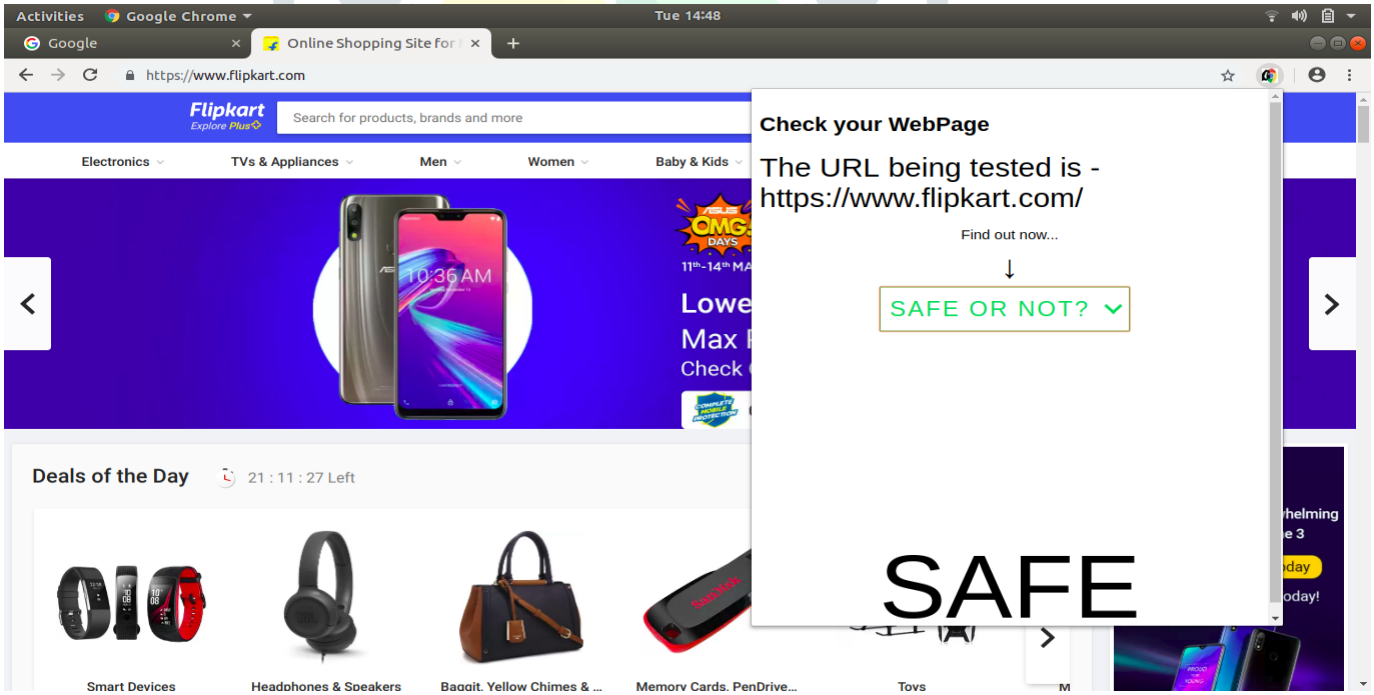


6.1 Data Flow Diagram

In fig 6.1 we have shown that first we will take the data from various websites and will apply machine learning algorithms on that data. Now we will take candidates URL which he or she wants to surf and will verify it. If URL is found malicious user will get informed that the URL, he or she wants to visit not safe.

VII Result : Following are the few sample outputs of our project.





VIII Conclusion : After implementing our project successfully we can conclude that our system includes large scale experimental study of detecting malicious URLs, which includes massive set of URLs. We have almost collected nearly 4,00,000 URLs that contains 40% of malicious URLs. Our collection is not limited to particular service domain, but our collection also includes URLs from various domain like Bit.ly , Ow.ly, t.co ,goo.gl etc.

REFERENCES

1. Luca Invernizzi.Stefano Benvenuti[7] “Evilseed: A guided approach for finding malicious Webpage,” in IEEE Security Symposium, 2012.
2. M. Rajesh et al.[1] M. Rajesh, R. Abhilash and R. Praveen Kumar, "URL ATTACKS: Classification of URLs via Analysis and Learning", International Journal of Electricaland Computer Engineering, Vol. 6, No. 3, pp. 980- 985, 2016
3. Roshani et al. [4] Roshani K. Chaudhari and D. M. Dakhane , “Machine Learning Approach for Detection of Malicious Urls and Spam in Social Network”, International Research Journal of Engineering and Technology, vol. 03 , no.05, 2016
4. Bargal Varsharani et al.[3] Barhate Apeksha S, Bargal Varsharani D, Suralkar Rupali, Shewale Rekha V “Detecting Suspicious URLs using Bayesian Classification in OSN Data”, International Journal of Advance Research and Innovative Ideas in Education, vol-2, no.2, 2016.
5. DJ Guan, Chia-Mei Chen, and Qun-Kai Su.[5] Feature set identification for detecting suspicious urls using bayesian classification in social networks.