International Journal of Engineering, Science and Mathematics
Vol. 8 Issue 9, September 2019,
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@gmail.comDouble-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

DETECTION OF MALICIOUS WEB PAGES USING NAÏVE BAYES CLASSIFICATION AS OPPOSED TO K-MEANS TECHNIQUE

SPREEHA DUTTA

KEYWORDS:

Malicious web pages; Naïve Bayes;
K-Means;
Data mining

ABSTRACT

The main aim behind this project is to detect malicious web pages. So our approach is to detect URLs instead of domain names since if we blacklist a domain name, then even after a harmful link has been removed from it, the domain names stays blacklisted but if we run our detection for every individual URL on the web page, it removes the aforementioned risk.

From the attacker's point of view:

Most of the attackers inject a malicious URL in the starting page of a website since it is the starting page that hosts users to EKs via redirects. What attackers most commonly do is that they implant a **malicious anchor tag** (<a>).

What we will be doing:

First we will be crawling the web by using a starting URL of a page and thereby extracting all URLs and links present on that page with the <a> tag and href attribute. We will be securely storing all the extracted URLs into a database .

Each URL fetched from the database will then be analysed on a number of parameters like **length** of the URL, **frequency** of occurrence of characters in the URL and then they will be matched against the testing data with **malicious keywords** to train the URLs further. Based on these parameters we will adopt a decision tree approach and use **Naïve Bayes** classification to classify them as blacklisted or safe.

Naïve bayes is a technique that is mainly used for **classification**. All naive Bayes classifiers go with the basic assumption that the value of a particular feature is independent of the value of any other feature, when given with the class variable. The features like length , character frequency and their effect behind an URL being blacklisted

or not and whether these features are dependent on each other or not will be studies using this algorithm to arrive at meaningful **decisions**.

Finally, the results obtained will be evaluated with real data and the accuracy of our analysis will be measured.

Why are we using Naïve Bayes classification?

As clearly depicted in the graph below , Naïve Bayes has the highest true positive rate in comparison to other algorithms like support vector machine and neural network. Naïve Bayes gives the guarantee of maximum accuracy with ease of analysis of independence between different features which is why we decided to go along with this approach.

Author correspondence:

Software Engineer,
Bengaluru,India.

1. INTRODUCTION

Online Social Network, for example, Twitter enables its clients to, in addition to other things, small scale blog their everyday movement and discussion about their interests by posting short messages called tweets which are comprise of 140 characters. Twitter is very well known with in excess of 100 million dynamic clients who post around 200 million tweets each day . As the dispersal of data is exceptionally simple on Twitter, makes it a well known approach to spread outer substance like articles, pictures and recordings by implanting URLs in tweets. Be that as it may, these URLs may connection to low quality substance, for example, malware, spam sites or phishing sites. Malware, short for noxious programming, is programming used to upset PC activity, gather delicate and vital data, or access private PC frameworks. Phishing is the demonstration to endeavor for getting data, for example, usernames, passwords, and Visa subtleties and now and then, in a roundabout way cash by taking on the appearance of a solid element in an electronic correspondence. Spam is flooding the Internet with various duplicates of a similar message, in an undertaking to constrain the message on client or individuals who might not generally get it.

Facebook is inclined to malignant posts having spam URLs for dissemination. Customary facebook spam discovery techniques exploit account highlights, for example, the proportion of posts

containing URLs and the date of making a record, or connection includes in the facebook diagram. These recognition techniques are inadequate against highlight manufactures or devour much time and assets. In this report we have proposed an AI way to deal with find Malicious URLs and spam and to distinguish whether a given post is spam or not in a Social media, for example, Facebook. By gathering dataset and preparing the classifier we ordered the info post. The Naive Bayes calculation, a managed learning model with related learning calculations which are utilized to break down information utilized for arrangement and relapse investigation. After grouping the affectability of each post is determined. After trial results it is discovered that the prepared classifier is appeared to be precise and has low false positives and negatives.

2. PROBLEM STATEMENT

In all current Online Social Networks (OSNs) the client-server architecture is embraced. The OSN specialist service provider acts as the controlling entity. All the content in the framework are put away and overseen by it. OSN is utilizing on the web spam sifting is introduced at the OSN specialist organization side. Once introduced, it review separate message before perusing the message to the proposed beneficiaries and settles on critical choice on whether the message under examination ought to be dropped. In the event that the message is unlawful mean right away dropped the message else it is sent to the relating recipient. Diverse Twitter spam discovery plans have been proposed, to adapt to noxious tweets. These plans can be partitioned into record include based and connection highlight based plans.

Record highlight based plans utilize the separating highlights of spam records, for example, the proportion of tweets containing URLs, the date of record creation, and the quantity of supporters and companions. Be that as it may, noxious clients can without much of a stretch think up these record highlights. The connection highlight put together plans depend with respect to progressively powerful highlights that malignant clients can only with significant effort gather, for example, the separation and availability clear in the Twitter diagram. Getting these connection highlights from the Twitter chart, notwithstanding, requires a vital measure of time and assets, on the grounds that the Twitter diagram is marvelous in size. Numerous suspicious URL location plans have additionally been presented. They utilize static or dynamic crawlers and might be executed in virtual machine honeypots, similar to Capture-HPC , HoneyMonkey, and Wepawet, to look at recently watched URLs. These plans isolate URLs as per a few highlights containing DNS data, lexical highlights of URLs, URL redirection, and the HTML substance of the points of arrival. Notwithstanding, pernicious servers can sidestep examination by specifically giving benevolent pages to crawlers.

3. OBJECTIVE

In this AI approach, a discovery of noxious URLs is finished utilizing the gathered dataset, instead of investigating the points of arrival of individual URLs , which may not be effectively gotten, we manage related divert chains of URLs . Since aggressors' assets are limited and should be reused, a piece of their divert chains must be shared. We found an alternate number of important highlights of suspicious URLs got from the corresponded URL divert chains and related setting data. We collected a Dataset which contains huge number of Malicious URLs tweets from the Kaggle and prepared a factual classifier with their highlights. From results it is discovered that the prepared classifier has high precision and low false-positive and false-negative rates.

When Naive Bayes classifier is used for learning from training dataset, it never contains more attributes than components available and it guarantee that assumption will be brought up effectively. In this paper represents experimental evaluation for classification and detection of URLs using Naïve Bayes and K means , whereas Naive Bayes classifier is basically a probabilistic classifier based on assumption. On the basis of assumption and learning from train set; it finds out most suitable assumption based on previous assumptions and initial knowledge.

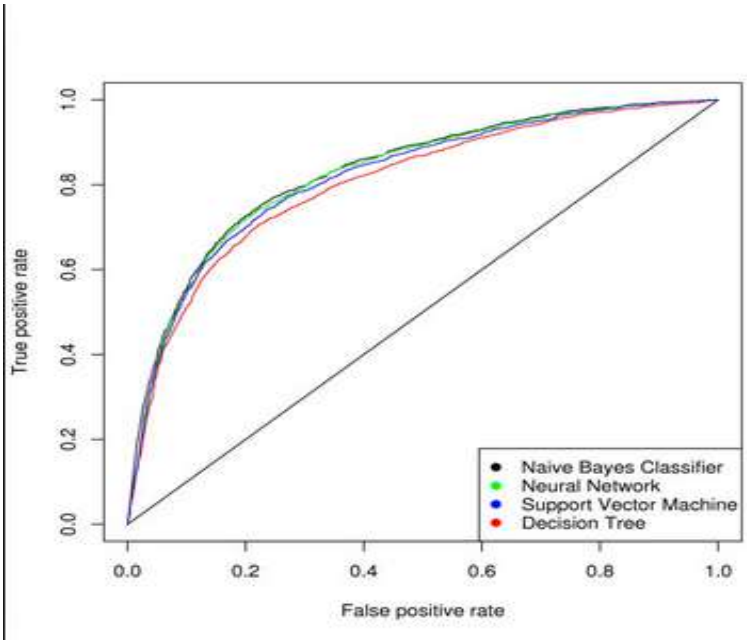


Fig 1: ROC curve depicting performance of different algorithms

4. LITERATURE SURVEY

In the recent times a lot of research work has been carried out for the design a better detection mechanism. G. Stringhini, G. Vigna and C. Kruegel in 2010 [4] used account features such as Friend-Follower ratio, URL ratio and message similarity to differentiate spam tweets. This paper resolves to which extent spam has entered social network and how spammers who points social networking sites operate. To assemble the data about spamming activity, a large and disparate set of “honey-profiles” are established on three large social networking sites and then analyzed the

collected data and identified peculiar behavior of users who influenced honey-profiles. Features are developed based on the analysis of this behavior which is used for detection. A. Wang in 2010 [5] modeled Twitter as directed graph where user accounts are represented by vertices and the type of relationship between users, friend or follower is actuated by the direction of edge. In this paper, detection mechanism is based on graph based features like in-degree and out-degree of nodes and content based features like presence of Trending topics and HTTP links in tweets. This work applies machine learning methods to automatically discriminate spam accounts from normal ones. Based on the API methods provided by Twitter to excerpt public available data on Twitter website, a Web crawler is developed. Finally, a system is established to assess the detection method. J. Song, S. Lee, and J. Kim in 2011[6] viewed Twitter as an undirected graph and made use of Menger's theorem to evaluate the values of message features such as distance and connectivity between nodes in order to achieve detection. The relation features prototype system such as distance and connectivity are exclusive features of social networks and are difficult for spammers to forge or manipulate. This system analyses spammers in real-time, this implicates that when a message is being delivered, clients can classify the messages as spam or benign. C. Yang, R. Harkreader, and G. Gu (2011) [7] in their research used time based aspects such as tweet rate and following rate besides graph based aspects and content based aspects in order to perform detection. H. Gao, Y. Chen, K. Lee, D. Palsetia, and A. Choudhary [8] suggested a detection system based on message features such as interaction history between users, average number of tweets containing URL, average tweet rate, and unique URL number. In OSNs, multiple users are connecting and interacting via the message posting and viewing interface. The system analyses every message and calculates the feature values before rendering the message to the intended recipients and makes immediate determination on whether or not the message under investigation are dropped.

5. METHODS

5.1 Existing Methods

Knowing the kind of attack or threat that we might be facing enables us to estimate the severity of the attack and thereby aids in adopting an effective countermeasure. Existing methods generally detect malicious URLs of a single attack type. They consider only one of the attributes among many like link structure, length, character frequency and do not take into account all of the attributes that one needs to consider in order to precisely decide the maliciousness of an URL. There are several batch algorithms available that process batches or large sets of data and there are online algorithms available too to analyse data in real time. There are a few efficient heuristics and determining rules to differentiate between safe websites from the ones that are dangerous. The internet criminals rely on the absence of these heuristics to set their targets.

One of the most common techniques that is deployed in browser toolbars and filtering appliances that filter the web like search engines is by using blacklisting. In this methodology, an outsider

administration aggregates the names of all common bad web destinations (marked by mixes of client input, Web slithering, and heuristic investigation of webpage content) and circulates the rundown to its supporters. While such frameworks have insignificant query overhead (just searching for a URL within the list) they can only offer incomplete security in light of the fact that no blacklist is far reaching and undesirable. Along these lines, a client may tap on a malicious URL before it shows up on a blacklist.

Then again, a few web intercepts likewise block and investigate each website's full content as it gets downloaded. In spite of the fact that this investigation can recognize bothersome sites with higher exactness, it acquires by far more runtime overhead than questioning a blacklist; it might likewise accidentally open clients to the very dangers they look to stay away from.

5.2 Proposed Method

We will be detecting malicious URLs from a dataset taken from Kaggle and be analysing it on whether the URLs are safe or not based on two algorithms. We will be using Naive Bayes algorithm to determine the accuracy with which we classify the training data against the testing data. We will be seeing the relation, dependence or independence of attributes like length and rank of the URL. We will be using K-Means algorithm as the second method to classify the URLs into clusters.

Naive Bayes:

Naïve Bayes when compared with other algorithms has the highest true positive rate than others like support vector machine(SVM) and neural network. Naïve Bayes gives the guarantee of the maximum amount of accuracy along with ease of analysis of independence between different features which is why we decided to go along with this approach.

Naive Bayes comes from a family of probabilistic algorithms that take merit of the existing probability theory and also Bayes' Theorem to predict the tag of a text (since often attackers implant malicious tags in the anchor tag of sites). They are probabilistic in nature, which means that they are used for calculating the probability of each tag for text enclosed within a tag, and then output the tag with the highest one. That way naive Bayes algorithm collects these probabilities with the help of Bayes' Theorem that which clearly enunciates the probability of a feature based on some prior knowledge of conditions that may be related to that feature under study.

K-Means Clustering:

K-means clustering is one of the most popular unsupervised machine learning algorithms.

The K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids.

It halts creating and optimizing clusters when either:

- The centroids have stabilized—there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved.

K means is extremely popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. For each seed, there is a corresponding set of points encompassing the region that are nearer to the seed.

6. EXPERIMENTAL SETUP AND RESULTS

Here we are using Python programming language that is integrated on Spyder. Spyder is a powerful scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts. It offers a unique combination of the advanced editing, analysis, debugging, and profiling functionality of a comprehensive development tool with the data exploration, interactive execution, deep inspection, and beautiful visualization capabilities of a scientific package.

Beyond its many built-in features, its abilities can be extended even further via its plugin system and API. Furthermore, Spyder can also be used as a PyQt5 extension library, allowing you to build upon its functionality and embed its components, such as the interactive console, in your own software.

Libraries like pandas, numpy, kMeans, sklearn have been imported for this project.

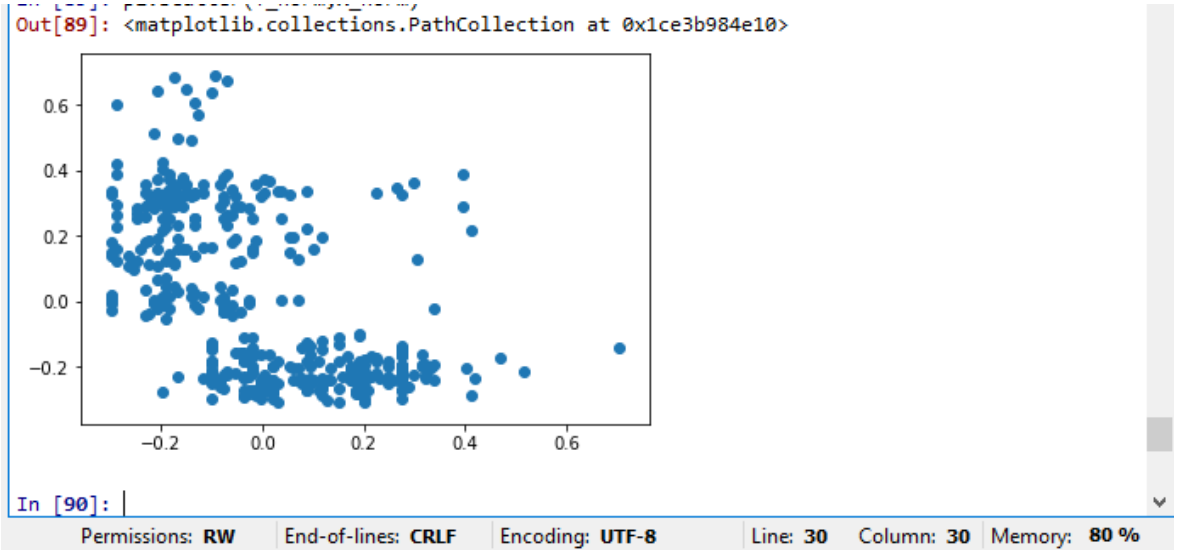
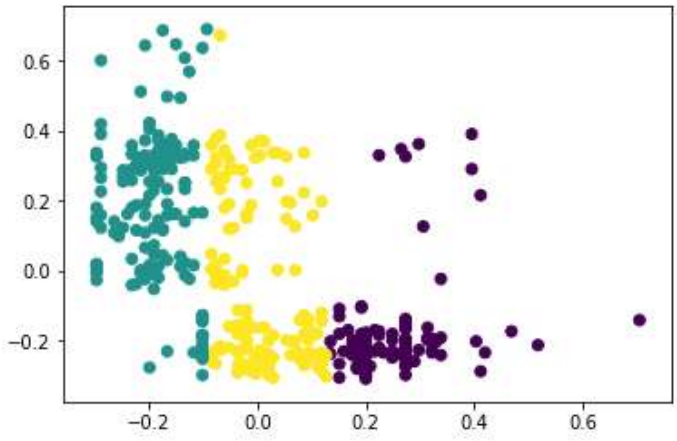


Fig 2: Cluster formation- length vs rank of URL

```
In [112]: pl.figure('3 Cluster K-Means')
Out[112]: <Figure size 432x288 with 0 Axes><Figure size 432x288 with 0 Axes>

In [113]: pl.scatter(pca_d[:, 0], pca_c[:, 0], c=kmeansoutput.labels_)
Out[113]: <matplotlib.collections.PathCollection at 0x1ce3ba4e940>
```




```
In [114]:
```

Fig 3: Final cluster formation

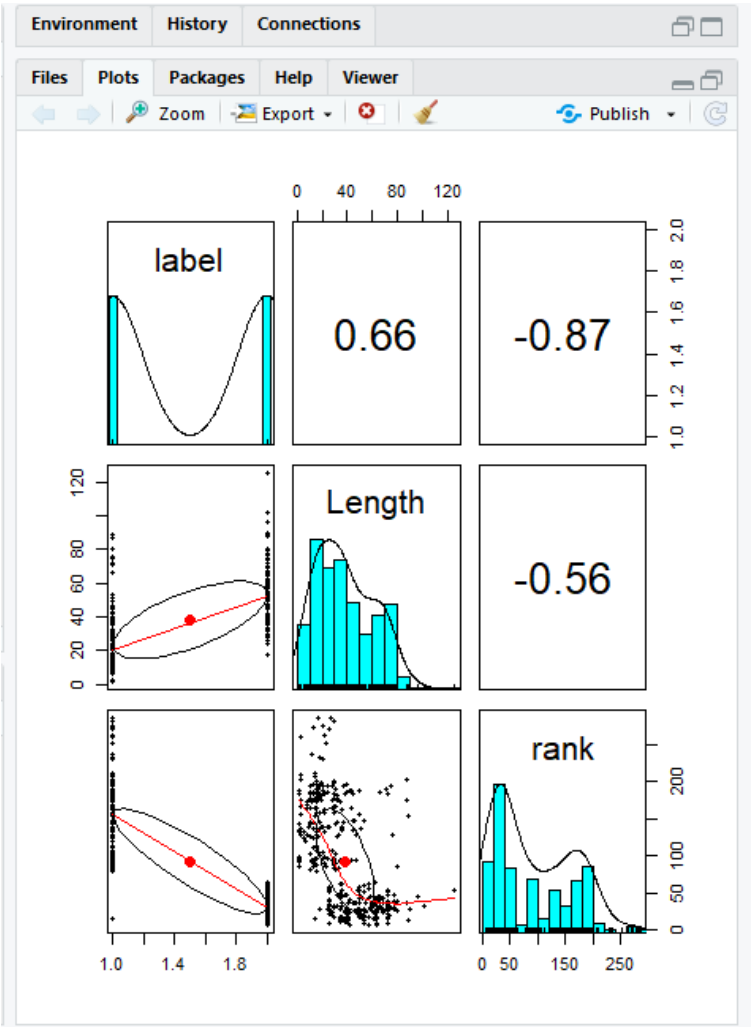


Fig 4: Figures depicting the cluster coefficient and correlation between attributes

7. CONCLUSION

Naive Bayes was used since it is said to give more accurate results than its other counterpart algorithms. It yielded an accuracy rate of 97.53% in classifying the training dataset on the basis of whether the URLs are malicious or not. Mean and standard deviation and graphs that plot the correlation between the URL attributes of length and ranking in search results depict that the URLS with length more than 90 have a much higher chance of being malicious. The top ranked search results have a lesser chance of being malicious than the ones that have a rank of more than 50 and feature on the fourth and fifth page of the search results. We found naive Bayes to be a more accurate and better classification technique than Kmeans since KMeans has a disadvantage that the number of clutsers need to be determined beforehand and the accuracy rate given by Naive Bayes was also more than that of KMeans.

REFERENCES.

[1] [1] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In Int. World Wide WebConf. (WWW), 2010.

[2] [2] THOMAS, K., GRIER, C., MA, J., PAXSON, V., AND SONG, D. Design and Evaluation of a Real-Time URL Spam Filtering Service. In Proceedings of the IEEE Symposium on Security and Privacy (May 2011).

[3] [3] ANDERSON, D. S., FLEIZACH, C., SAVAGE, S., AND VOELKER, G. M. Spamscatter: characterizing internet scam hosting infrastructure. In Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium (Berkeley, CA, USA, 2007), USENIX Association, pp. 10:1–10:14.

[4] [4] G. . Stringhini, C. Kruegel, and G. Vigna, “Detecting Spammers on Social Networks,”Proc. 26th Ann. Computer Security Applications Conf. (ACSAC), 2010.

[5] [5] A. Wang, “Don’t Follow Me: Spam Detecting in Twitter,” Proc. Int’l Conf. Securityand Cryptography (SECRYPT), 2010.

[6] [6] J. Song, S. Lee, and J. Kim, “Spam Filtering in Twitter Using Sender-ReceiverRelationship,”Proc.14th Int’l Symp. Recent Advances in Intrusion Detection.(RAID), 2011.

[7] [7] C. Yang, R. Harkreader, and G. Gu, “Die Free or Live Hard? Empirical Evaluationand New Design for Fighting Evolving Twitter Spammers,” Proc.14th Int’l Symp.Recent Advances in Intrusion Detection (RAID), 2011.

[8] [8] H. Gao, Y. Chen, K. Lee, D. Palsetia, and A. Choudhary, “Towards Online SpamFiltering in Social Networks,” Proc. 19th Network and Distributed System SecuritySymp. (NDSS), 2012.

[9] [9] Harry Zhang "The Optimality of Naive Bayes". FLAIRS 2004 conference.

[10] [10] Caruana, R. and Niculescu-Mizil, A.: "An empirical comparison of supervised learning algorithms". Proceedings of the 23rd international conference on Machine learning, 2006.