www.revistageintec.net
ISSN: 2237-0722
revista GEINTEC Gestão, Inovação e Tecnologias

SQL Injection Detection Using Machine Learning

S.S. Anandha Krishnan[1]; Adhil N Sabu[2]; Priya P Sajan[3]; A.L. Sreedeep[4]

[1]Noorul Islam Centre for Higher Education, Bachelor of Technology-IT with Specialization in Cyber Security and Forensics, India.

[1]anandhansuneev0704@gmail.com

[2]Noorul Islam Centre for Higher Education, Bachelor of Technology-IT with Specialization in Cyber Security and Forensics, India.

[2]adhilsabu1010@gmail.com

[3]Project Engineer, C-DAC Thiruvananthapuram, India.

[3]priyasajan@cdac.in

[4]Project Engineer, C-DAC Thiruvananthapuram, India.

[4]sreedeep@cdac.in

Abstract

SQL Injection attack considered as the one among the foremost vulnerability that exploits in terms of privacy exposure also as money loss. SQL Injection attacks are the amount one vulnerability within the most up-to-date OWASP top 10 report, and therefore the amount of those attacks continues to extend. Machine learning algorithms are used to solve the SQL Injection detection challenge. The next traffic is classified as SQL Injection or plain text using a classification algorithm. The machine learning classification algorithm is used in the matter, they are, Naive Bayes Classifier, Passive Aggressive Classifier, SVM, CNN, Logistic Regression. As a result, CNN was chosen to be used in the detection of SQL Injection attacks.

Key-words: SQL-injection, Machine Learning, Supervised Learning, Online Learning, Naïve Bayes Classifier, Passive Aggressive Classifier, SVM, CNN, Logistic Regression.

1. Introduction

In a day most of the application we are using is internet based. Institutions prefer to create the web based applications to extend the exposure they acquire. Being exposed to Internet has also lead to a raise in the amount and severity of the web attacks. As the Internet has grown in popularity, we have become accustomed to conducting a variety of transactions online. SQL Injection Attack is a

type of internet intrusion tactic that uses SQL to perform attacks on databases and falsify it for the purpose of determining customer neediness. SQL Injection Attacks became an accelerative explanation to concern by cyber protectors. Recently, utilization of machine learning algorithms to discover or stop different cyber safety issues are actually considered mostly. At the same time facility of applying supervised and online learning methods for discover security threats can't be interrogated, the computing informant and period needed for run those complicated algorithms remain a serious interest for the ever progressive cyber security organizations. Here an entry to SQL Injection attacks and thus the necessity along with motive towards make an improved system for SQL Injection detection. To know the SQL Injection attacks and different kinds intimately in later parts. For every executions and study finished thus supply sufficient literature review to find out here and change on the matter. Supervised learning and online learning, is the general formulation we are employed to unravel this issue.

2. SQL Injection

SQL injection, also known as SQLI, is a common attack vector that uses malicious SQL code for manipulating the backend database to access the data that was not intended to be expose. By using an SQL Injection vulnerability, given the proper circumstances, an intruder can use it to bypass an internet application's authentication and authorization mechanisms and retrieve the contents of a whole database.

Fig. 1 - The SQL Injection Attacks

SQL Injection attacks are often classified consistent with different criteria, attacker's motive (extracting data, determining database schema, inserting data, evading detection, implement denial of service, perform remote commands) and/or technical methods (Tautologies, Illegal/Logically Incorrect Queries, Union Query, Piggy-backed Queries, Stored Procedure, Alternate Encodings, Blind Injection, Timing Attacks)

A. Kinds of SQL Injection

1. Union Based SQL Injection
2. Error Based SQL Injection
3. Blind SQL Injection

1. Union Based SQL Injection

In union query attack, the attacker uses the UNION operator to hitch a malicious query to the first query. The results of the malicious query are going to be connecting the results of first query, allowing the attacker to get the values of columns of other tables.

2. Error Based SQL Injection

An error-based SQL Injection attack is carried out by sending invalid data into a query, which causes the database to make a mistake. This is frequently accomplished by forcing the database to do an operation that may result in a mistake. The user can then search for mistakes made by the database and use these issues to discover facts on how to advance in the database using SQL queries.

3. Blind SQL Injection

Blind SQL injection may be one sort of SQLI attack that generate true or false questions to the database and find out the solution supported the application's reply. Above Mentioned attack is usually applied at the online practice is aligned to point out general mistake alerts, but has not mitigated the code that is susceptible to SQL injection.

3. Related Work

William [1] introduced AMNESIA tool, a way to detect SQLI attacks before being executed on the database. It works supported static and run time investigation method to confirm the rightness of SQL queries. Valeur et al. [2] introduced another way for spot queries that didn't fit manifold models of regular queries at dynamic, consider string model and data type-autonomous model. However this didn't attain high accuracy, the tactic identifies broad possibility of deep learning in malicious query discovery. Gould et al. [3] formulated a JDBC checker. It is a static investigating

way to ascertain for wrongness in SQL strings and confirm them for possible malevolent queries. This checks the SQL strings for rightness and hope to spot and shows potential mistakes in SQL queries.

As against supervised learning, Bockermann et al. (2009) [4] introduced a model for using clustering (”internal self organizing maps”) SQL statements to parse tree construction of SQL queries as property. Y. Kosuga et al. [5] develop a system called as SANIA to detecting SQL injection exposure in web application at this event and correcting phase. Ke Wei, M. Muthuprasanna, Suraj Kothari[6] also provided a totally unique way to guard the stored process from attack and discover SQL injection. This approach joins dynamic check with static application code investigation in order that it will get rid of weakness to attack. For valuation and confirmation Pan et al. (2018) [7] introduced how to sort the practicability of using machine learning approach for web application breach detection, defined limits and challenges for this machine learning approaches, also as represented RSMT, one tool it utilise DNN and auto encoders for semi-supervised and unsupervised learning for web attack spotting, consider SQL injection. A.Ciampa et al (2010) [8] introduce a way to detect and stop SQL injection attack. That system predicated on examine web application to collect data like its construction and page form. Then it attempt to insert SQL code on the input area to form the online application shows a fault message. Then, it compares the output make by a web application and the structure of every aspect that same as to error message.

4. Machine Learning

Machine learning is a method of computer algorithms that mend mechanically via experience. Machine learning algorithms create a model that supports example data, mentioned as "training data", so as to form prevision or judgment requiring being explicitly programmed to attempt to so Machine learning algorithms are used in a great variety of uses, same as email filtering and computer vision, where it's hard or infeasible to create traditional algorithms to execute the required work.

There are 2 types of machine learning approach is used.

A. Supervised Learning

Supervised learning is a machine learning methodology that, in its most basic form, operates in the manner described. We have a dataset called training dataset, and each and every element of it is tagged. The supervised learning framework fundamentally finds the relationship between the data and

therefore the tag, and then uses this learned data to sort raw data that it has never seen before. The accuracy of a supervised learning algorithm is determined using a test dataset. Using supervised machine learning, this is frequently how we calculate values or classify never-before-seen data. Regression and classification algorithms are two types of supervised learning algorithms.

B. Online Learning

Online machine learning is a type of machine learning in which data becomes available in stages and is used to renew the simplest model for new data at each step, as opposed to batch learning, which produces the simplest predictor by learning on the entire training data set at once. Online learning is a common strategy used in the field of machine learning when it's impractical to train the entire dataset, necessitating the use of out-of-core techniques. It's always utilized in situations when the algorithm must dynamically adapt to new forms within the data, or when the data is generated as a function of time, such as stock price prediction. Online learning algorithms may also be vulnerable to detrimental interference, a drawback that incremental learning approaches will identify.

C. Machine Learning Algorithms

1. Naïve Bayes Classifier

Naive Bayes algorithm is a supervised learning method that is supported on Bayes theorem. This algorithms has already been enforced in SQL injection detection field. The core to Naïve Bayes is that it supposition of independency between all couple of property. It requires a little quantity of training data and extremely fast compared with other.

2. SVM

Support Vector Machine or SVM is one among the foremost favorite supervised learning algorithms, that is utilised for Classification as well as Regression problems. The aim of the SVM algorithm is to make the simplest line or decision boundary which will segregate n dimensional space into classes in order that we will well set the new datum within the exact family within early. The foremost decision boundary is named a hyperplane. SVM select the utmost points/vectors that used to

create the hyperplane. Those utmost cases are called as support vectors, and therefore algorithm is known as Support Vector Machine.

3. CNN

A convolutional neural network has numerous inputs and outputs since it is made up of many perceptrons. A line or an encrypt unit of signs can be fed into the neural network, which will then produce a set of judgments based on the correlation method. The term "convolutional neural network" refers to the system's use of a numerical process known as convolution. Convolutional networks are a special type of neural networks that use convolution in place of generic matrix multiplication in leastways one of it's layers.

1. Logistic Regression

Logistic regression is one among the for most touristed Machine Learning algorithms, which coming below the Supervised Learning method. It is put-upon to predict the categorical dependent variable using a given set of autonomous variables. Logistic regression is utilized for resolve the classification trouble. Logistic regression is a classification algorithm, utilized to guess the chance of happening of an event (0/1, True/False, Yes/No). It uses sigmoid function to figuring chance.

2. Passive Aggressive Classifier

Passive Aggressive Classifier is one of the few 'online-learning algorithms. This algorithm is perfect for sort out large streams of data. It is simple to apply and very fast. It works by fetching a model, learning from it and then working by the experience from the example. This algorithm have two parts Passive and Aggressive.

Passive: If the prediction is correct, the algorithm doesn't make any modification to the model.

Aggressive: If the prediction is incorrect, the algorithm make modifications to the model.

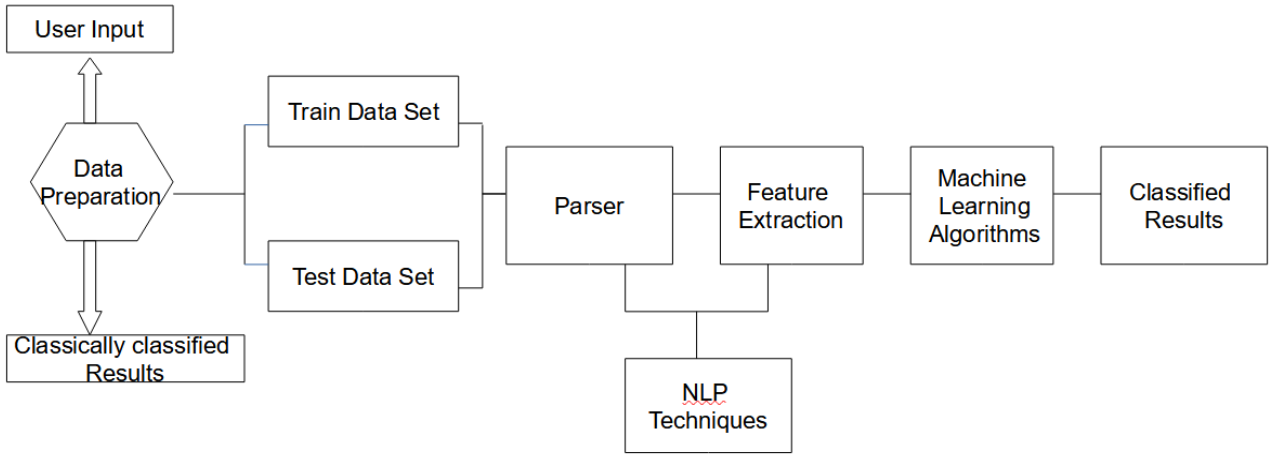
5. Proposed Work

Having the best skilled and the most well-guarded applications are always at odds with one another. There is no way to achieve complete safety without some investment. In this case, the

expenditures usually outnumber the time allotted for performance. The majority of apps vulnerable to SQL injection attacks are web applications. Here are several famous types of user-created SQL inputs that academics are experimenting with, as well as various methodologies and tools for detecting and avoiding such harmful inputs. When it comes to SQL injection attacks and application security, there are a number of challenges and issues that we face.

In this area, we projected a unique framework using machine learning to detect SQLi attacks. Instead of tackling the injection on the injection attacks on the server side by guarding the database, this model is designed to act as a filter on the client side.

Fig. 2 - Block Diagram for Proposed Model



A. Dataset Pre-processing

Pre-processing of data means change the dataset to lowercase, avoid unwanted spacing or blanck space, etc. The disputation of clearing covariant and merit from statements like 1=1 and 2=0, that avoid mathematical statement symbol and change some possible trigger states in first approaching, which we later derelict in approval of the adjacent one, was only supported on dismissal of entropy from the trace. All the left functions, escape symbols, and statements where then coded as a single value.

Data Preparation: For the purpose of reducing data noise and improving precision, unnecessary spaces and escape sequences were eliminated and all the queries were converted into lowercase form.

Training and Testing Datasets: The training and testing sets are randomly taken from the dataset with a conventional ratio of 80-20 (80% for training and 20% for testing) usingtrain_test_splitfunction built into sklearn library:

train_test_split(trainDF['text'], trainDF['label'], test_size = 0.2)

Parser: Our model came across a common adversity during the data processing phase where traditional machine learning model explicitly takes in structured tabular numeric data, but our collected data are entirely non-structured texts. This is where parser for text comes into play. Text parsing is the process of transforming given series of text into desired smaller components based on some specific rules. There are two common ways to parse texts: regular expression separation and tokenization. The former parses the targeted text using desired regular expressions such as "[a-z]", "[\t]".The latter divides the text into tokens, where each token can be a character, a word or a phrase. In the case of SQLI attacks, the regular expressions do not determine the malice of a query, the appropriate text parsing method for this model is tokenization. Queries are split into tokens of words. For example: Parsing "or 1=1 -- 1" into "or", "1=1", "--", "1".

B. Machine Learning Approach

Supervised learning and online learning methodology was chosen for the present work as the best method to the trouble, that has some active models (SQL injection and XSS) and need a non-linear judgment-making.

Here we are using 5 different machine learning algorithms to detect best among them, They are Naive Bayse, Logistic Regression, SVM, CNN, Passive Aggressive Classifier.

Natural Language Processing (NLP) techniques and feature extraction: For NLP, there are many featured engineering techniques, but the one that proves to be the most useful for this SQLi attacks detecting model is Word Level TF-IDF Vectors. TF-IDF stands for Term Frequency and Inverse Document Frequency, which is an important index for term searching and figuring out the relevancy of specific terms in a document. Term Frequency specifically compute how often a word occurs in a document, where Document Frequency determines how often a word happen in an entire set of documents. The most significant advantage of TF-IDF is that it will assume that the documents are just bags of words, where each word does not have any correlation to another. This method is simple but powerful for our use case since in SQL, there are no tense or grammar rules like human languages.

C. Optimisation

Optimisation is an important aspect of our strategy. When neural models of various types are compared in the same state (same dataset, same amount of training cycles, etc.), they show a spotting range of up to 97 percent. Manually selecting the best neural network architecture and training settings is a stochastic process that takes a long time to complete and requires re-evaluation of the work based on a variety of factors. Optimisation allows you to adjust those activities and select a broad execution configuration of elements without having to re-configure everything manually.

6. Evaluation Methodologies

A. Datasets

For the evaluation purpose we are using a dataset for training our model. We are collecting this dataset from github. From this dataset we are find the accuracy of each algorithm and find best model. So it is a necessary evaluation methodology for this work.

B. Metrics

The foremost possibility is then assess with the test dataset, and the factors of the algorithms that we were used to compare. This result is obtained from calculating True Positives (T P), True Negatives (T N), False Positives (F P), and False Negatives (F N). From this values, the system calculates accuracy, precision, and recall at the last of the valuation procedure. This may be represented in a matrix form and known as confusion matrix. By checking the FN and FP values we decide which parameter is used to evaluate.

7. Results

Researches were conducted using the above algorithms, using Vector feature Tf-Idf vectors at Word level and Ngram-level. Accuracy were recognized for all models. Here used K-fold cross validation method to enhance the potency of the models. For this experiment, we applied text classification on the available datasets from website named kaggle. From this classification we got

CNN is that the best algorithm for future work and it'll saved for detect SQL Injection attacks. The CNN algorithm shows 97% accuracy with this model.

Table 1 - Parameter Score of ML Algorithms

ALGORITHM	ACCURACY	PRECISION	RECALL
Naive Bayse	95	0.85	0.98
Logistic Regression	92	0.97	0.76
CNN	97	0.92	0.96
SVM	79	1.0	0.20
Passive Aggressive	79	1.0	0.20

8. Conclusion

An SQL injection attack on a web application could be a major problem. It's critical to figure out a viable solution to this problem. Researchers have devised a number of methods for detecting and counteracting this flaw. There is no strategy that will prevent all types of SQL injection attacks. For cyber security professionals, SQL Injection attacks are still a key cause of concern. Signature-based SQL Injection detection systems are no longer reliable since attackers deploy new types of SQL Injections on a regular basis. SQL Injection detection systems must be able to distinguish between innovative, never-before-seen attacks. Machine learning is being considered by many researchers for use in the realm of cyber-security. Because machine learning in cyber-security is still a growing field, there are numerous machine learning-specific libraries and open source tools that can be utilised to solve problems with threats and attacks.

The SQL Injection Detection Challenge is solved using machine learning algorithms. A classification method is utilized to classify incoming communication as SQL Injection or normal text. Five machine learning techniques are used to classify the problem: Nave Bayes Classifier, Passive Aggressive Classifier, SVM, CNN, and Logistic Regression. The Nave Bayes classifier machine learning model has a 95 percent accuracy rate, while the Passive Aggressive Classifier has a 79 percent accuracy rate, SVM has a 79 percent accuracy rate, and Logistic Regression has a 92 percent accuracy rate. Because they use numerous basic classifiers to improve error and accuracy, supervised learning methods are considered to provide results with finer accuracy. As a result, out of all the algorithms examined, CNN was chosen to be used to the SQL Injection categorization problem. The CNN algorithm achieves a 97 percent accuracy by tuning and trying a variety of parameters simultaneously.

9. Future Scope

As part of a future study, we'll be measuring numerous web-based application codes in the public domain in order to achieve high accuracy in SQL injection detection methods. SQL integration with nikto HTTP scanner, HTTP scanning proxies, and metasploit will aid in the detection of many web-based threats. Here CNN is used for future studies and used to create detection model, for better model we can use another algorithms and find best model. Static analysis method is used here we can add run time analysis as a future work. Here only SQL Injection detection system is created we can create a prevention system also in future.

References

William G.J. Halfond, Alessandro Orso, "AMNESIA: analysis and monitoring for NEutralizing SQL-injection attacks", *Proceedings of the 20th IEEE/ACM international Conference on Automated software engineering,* November 07-11, 2005.

F. Valeur, D. Mutz, and G. Vigna, *"A Learning-Based Approach to the Detection of SQL Attacks,"* 123–140, 2005.

C. Gould, Zhendong Su and P. Devanbu, "JDBC checker: a static analysis tool for SQL/JDBC applications," *Proceedings. 26th International Conference on Software Engineering,* Edinburgh, UK, 2004, 697-698.

C. Bockermann, M. Apel, and M. Meier, "Learning SQL for database intrusion detection using context-sensitive modelling (extended abstract)," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 5587 LNCS, 196–205, 2009.

Y. Kosuga, K. Kono, M. Hanaoka, ―Sania: Syntactic and Semantic Analysis for Automated Testing against SQL Injection. *The 23rd Annual Computer Security Applications Conference,* 107- 116.

Wei, K., Muthuprasanna, M., & Suraj Kothari. (2006). Preventing SQL injection attacks in stored procedures. *Software Engineering IEEE Conference.* 2007. http://ieeexplore.ieee.org

Y. Pan, F. Sun, J. White, D.C. Schmidt, J. Staples, and L. Krause, "Detecting Web Attacks with End-to-End Deep Learning," *Acm*, 1–14, 2019. https://www.dre.vanderbilt.edu/{∼} schmidt/PDF/machine-learning-feasibility-study.pdf

A. Ciampa, C.A. Visaggio, M.D. Penta, ―A heuristic-based approach for detecting SQL-injection‖, *Proceedings of the 2010 ICSE,* 2010.