People's Democratic Republic of Algeria

Ministery of Higher Education and Scientific Research

Ferhat Abbas University of Setif 1

Université Ferhat Abbas Sétif 1

Faculty of Sciences

Computer Science Department

DISSERTATION

Presented in fulfillment of the requirements of obtaining the degree

Master 2 in Computer Science

Specialty: Data Engineering and Web Technologies

THEME

Detecting SQL injections using Deep Learning techniques

Presented by:

ATHMANI RAMI

BOUHEZILA NASSIM

Supervised by:

Dr. BENZINE MEHDI

2022/2023

Dedication

To our parents,

To our grandparents,

To our brothers and sisters,

To our entire family,

To all our friends.

ATHMANI Rami,
BOUHEZILA Nassim.

Abstract

Deep learning techniques have improved various domains by using their ability to learn complex patterns from large datasets. In this dissertation, we employed the power of Deep Learning, specifically BERT language model (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers), to resolve the issue of SQL injection attacks on web applications.

The goal of our study is to develop a Deep Learning model using BERT that can accurately identify SQL injections.

Based on the results, our model demonstrated excellent performance; it also indicated that BERT outperforms the compared machine learning models across different evaluation metrics. These results affirm the effectiveness of BERT in detecting SQL injection attacks, underscoring its superior performance in our study.

Keywords: Deep learning, Deep learning techniques, Deep Learning model, BERT language model, SQL injection attacks, Web applications, Machine learning models, Evaluation metrics.

Résumé

Les techniques d'apprentissage profond ont amélioré divers domaines en utilisant leur capacité à apprendre des complexes patterns à partir de grands ensembles de données. Dans ce mémoire, nous avons utilisé la puissance de l'apprentissage profond, en particulier le modèle de langage BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers), pour résolu le problème des attaques par injection SQL sur les applications web.

L'objectif de notre étude est de développer un modèle d'apprentissage profond en utilisant BERT qui identifier les injections SQL avec précision.

D'après les résultats, notre modèle a démontré d'excellentes performances ; ils ont également indiqués que BERT a surpassé les autres modèles d'apprentissage automatique comparés à travers différents métriques d'évaluation. Ces résultats confirment l'efficacité de BERT dans la détection des attaques par injection SQL, confirmer sa performance supérieure dans notre étude.

Mots-clés: Apprentissage profond, Techniques d'apprentissage profond, Modèle d'apprentissage profond, Modèle de langage BERT, Attaques par injection SQL, Applications web, Modèles d'apprentissage automatique, Métriques d'évaluation.

ملخص

حسنت تقنيات التعلم العميق مجالات مختلفة باستخدام قدرتها على تعلم الأنماط المعقدة من مجموعات البيانات الكبيرة. في هذه الأطروحة، استخدمنا قوة التعلم العميق، وتحديداً نموذج اللغة "BERT" (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) ، لحل مشكلة هجمات حقن SQL على تطبيقات الويب.

الهدف من دراستنا هو تطوير نموذج التعلم العميق باستخدام BERT الذي يمكنه تحديد هجمات حقن SQL بدقة.

بناءً على النتائج المتحصل عليها، أظهر نموذجنا أداءً ممتازًا؛ كما أشار إلى أن BERT يتفوق في الأداء على نماذج التعلم الآلي التي تم المقارنَة بها عبر مقاييس التقييم المختلفة. تؤكد هذه النتائج فعالية BERT في اكتشاف هجمات حقن SQL ، مما يؤكد أدائها المتفوق في دراستنا.

الكلمات الدالة: التعلم العميق، تقنيات التعلم العميق، نموذج اللغة BERT، هجمات حقن SQL، تطبيقات الويب، نموذج التعلم العميق، نماذج التعلم الآلي، مقاييس التقييم.

Table of contents

List of figures

List of tables

General Introduction

The rapid growth of web applications has revolutionized the way we interact and conduct various activities online. From e-commerce platforms and social networks to financial systems and government portals, web applications have become an integral part of our daily lives. However, with greater dependence on online applications comes an increased danger of cyber attacks with SQL injections being one of the most common and dangerous vulnerabilities.

To mitigate the growing threat of SQL injections, traditional approaches such as input validation and query parameterization have been widely adopted. While these methods provide some level of protection, they often struggle to keep pace with the evolving attack techniques employed by adversaries. Thus, there is a pressing need for more advanced and proactive defense mechanisms to detect and prevent SQL injection attacks.

In recent years, deep learning approaches have emerged as a promising solution in various domains, leveraging their ability to automatically learn complex patterns from large datasets. One such powerful deep learning model is BERT (Bidirectional Encoder Representations from Transformers), originally developed for natural language processing tasks. BERT has proven to be highly effective in capturing the semantic and contextual understanding of text, leading to remarkable performance in tasks such as text classification.

In this research, we propose using the power of BERT-based deep learning models to address the critical issue of SQL injection attacks. Our objective is to develop a reliable and efficient detection model capable of accurately identifying SQL injection attempts in real-time. By using BERT's contextual understanding and semantic representation capabilities, we aim to create a model that can effectively distinguish between normal and SQL malicious queries.

Our thesis is organized as follows:

In Chapter 1, we explore SQL injection attacks, their definitions, types and their detecting techniques, then we head on machine learning and Deep Learning, we present popular algorithmic approaches in machine learning and explore deep learning architectures in Chapter 2. Chapter 3 is dedicated to the general conception of our work and the materials used,

including the dataset and the type of deep learning architecture employed. Furthermore, we cover the preprocessing steps taken to ensure the accuracy and efficiency of our system. In Chapter 4, we discuss the test and evaluation of our model for detecting SQL injection attacks. We use a variety of evaluation metrics, including accuracy, precision, recall, and F1 score. We also compare the performance of our model to other machine learning algorithms and related works.

Chapter 1

SQL Injections

1.1 Introduction

In today's interconnected digital landscape, web applications have become an integral part of our daily lives. They enable us to perform a wide range of tasks, from online shopping and banking to social networking and communication. However, the rise of web applications has also brought about a corresponding increase in security threats and vulnerabilities like gaining unauthorized access, stealing sensitive information, or disrupting services.

The following list shows the Top Five (05) most dangerous web application security risks published by OWASP Top 10 2021 project [1]:

Broken Access Control: refers to inadequate restrictions on what authenticated users can access or perform within an application. It can lead to unauthorized access, privilege escalation, or exposure of sensitive functionalities or data.

Cryptographic Failures: refer to vulnerabilities or weaknesses in the implementation or use of cryptographic techniques and algorithms. These failures can result in the compromise of encrypted data, unauthorized access, or the ability to tamper with cryptographic operations.

Injection Attacks: such as SQL, OS, or LDAP injection, occur when untrusted data is sent to an interpreter as part of a command or query, allowing an attacker to execute malicious code or unauthorized actions.

Insecure Design: refers to a fundamental flaw in the architecture or design of a system or application that compromises its security. It involves the failure to incorporate proper security mechanisms, access controls, or threat modeling during the design phase. **Security Misconfiguration:** results from insecure configuration settings, default passwords, open cloud storage, or verbose error messages. Attackers exploit these misconfigurations to gain unauthorized access, extract sensitive information, or launch further attacks.

Focusing on SQL injection attacks, the impact of them on web applications can be severe. As an example, the attackers may have the ability to take control of the entire website or steal sensitive data like usernames, passwords, and credit card numbers. Furthermore, these attacks may result in data loss, website failures, and reputational harm to a company.

1.2 Understanding how web applications work

Web applications are defined as features that are most frequently seen on websites, such as shopping carts, product search and filtering, instant messaging, and social network newsfeeds. They enable users to access sophisticated functionality without having to install or set up software. The web application and users data are stored in a system called DBMS *"Database Management System"*, web apps that use these databases are called Database-driven Web applications.

Database-driven Web applications are very common in today's Web-enabled society. They normally consist of a back-end database with Web pages that contain server-side scripts written in a programming language that is capable of extracting specific information from a database depending on various dynamic interactions with the user. A database-driven Web application commonly has three tiers: Presentation tier (a Web browser or rendering engine, HTML, CSS, JS), logic tier (a programming language, such as C#, PHP, JSP, JS, etc.), storage tier (a database such as Microsoft SQL Server, MySQL, Oracle, etc.).Figure 1.1 illustrates the simple three-tier example.

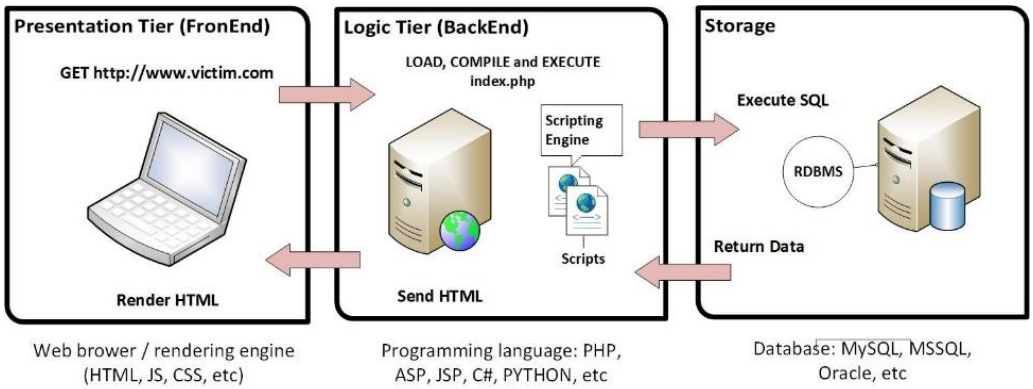

Figure 1.1 Three-tier architecture.

The Web browser (the presentation tier, such as Internet Explorer, Safari, Firefox, etc.) sends requests to the middle tier (the logic tier), which services the requests by making queries and updates against the database (the storage tier).

A fundamental rule in a three-tier architecture is that the presentation tier never communicates directly with the data tier; in a three-tier model, all communication must pass through the middleware tier. Conceptually, the three-tier architecture is linear.

In Figure 1.1, the user fires up his Web browser and connects to http://www.victim.com. The Web server that resides in the logic tier loads the script from the file system and passes it through its scripting engine, where it is parsed and executed. The script opens a connection to the storage tier using a database connector and executes an SQL statement against the database. The database returns the data to the database connector, which is passed to the scripting engine within the logic tier. The logic tier then implements any application or business logic rules before returning a Web page in HTML format to the user's Web browser within the presentation tier. The user's Web browser renders the HTML and presents the user with a graphical representation of the code. All of this happens in a matter of seconds and is transparent to the user.

1.3 How SQL injections work

1.3.1 Definition

SQL injection attack consists of the insertion or "injection" of a SQL query via the input data from the client to the application, they are introduced when software developers create dynamic database queries constructed with string concatenation, which includes user-supplied input. A successful SQL injection exploit can read sensitive data from the database, modify database data (Insert/Update/Delete), execute administration operations on the database (such as shutdown the DBMS), recover the content of a given file present on the DBMS file system and in some cases issue commands to the operating system [2].

Here are some facts regarding damages caused by SQL injection attacks:

1. In 2018, a SQL injection attack on a Canadian cryptocurrency exchange called MapleChange resulted in the loss of almost all of the company's cryptocurrency holdings, which were worth around 6 million $ [1].
2. In 2017, a SQL injection attack on Equifax, a major credit reporting agency in the United States, compromised the personal information of over 147 million people.

[1] MapleChange hack:
https://www.ccn.com/newsflash-canadian-bitcoin-exchange-hacked-says-all-funds-are-gone/
Accessed June 03, 2023

The breach resulted in a settlement of up to 700 million $ in damages, including compensation for victims and penalties [2].

3. In 2016, a SQL injection attack on the Philippine Commission on Elections exposed the personal information of over 55 million voters, including their names, addresses, and biometric data. The breach resulted in a class-action lawsuit and cost the commission over $2 million in damages [3].

In addition to financial damages, SQL injection attacks can also result in reputational damage, loss of customer trust, and legal liabilities. Companies that suffer from SQL injection attacks may face lawsuits, regulatory fines, and a loss of business due to damaged reputation.

Basically, SQL injection process is structured in Four (4) phases:

1. The attacker sends the malicious HTTP request to the web application
2. The malicious code concatenated with the developer's SQL statement
3. The system submits the SQL statement to the backend database.
4. The Database system returns the sensitive data for the Attacker to be exploited after.

As shown in Figure 1.2 describes a login by a malicious user exploiting SQL Injection **vulnerability**. The Administrator will be authenticated on the application after typing: employee id=112 and password=admin. The attacker is trying to find a SQL injection vulnerability to log in to the application by adding a tautology to the developer's SQL statement. [3]

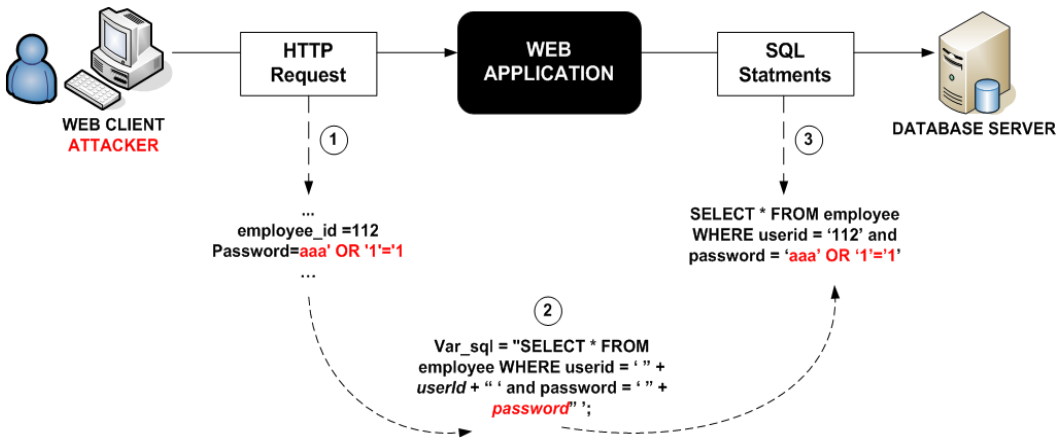

Figure 1.2 Example of a SQL injection attack [3].

[2] Equifax breach settlement: https://www.csoonline.com/article/3444488/equifax-data-breach-faq-what-happened-who-was-affected-what-was-the-impact.html - Retrieved June 03, 2023.

[3] Philippine Commission on Elections breach: https://www.reuters.com/article/us-philippines-election-idUSKCN0XI1FR - Accessed June 03, 2023.

The above SQL statement is always true because of the Boolean tautology that the attacker appended (OR 1=1) so, he will access the web application as an administrator without knowing the right password.

1.4 Techniques of SQL injections

1.4.1 Tautologies

This type of attack injects SQL tokens into the conditional query statement to be evaluated as always true, it is used to bypass authentication control and access to data by exploiting vulnerable input fields which use WHERE clause for example, consider the following SQL query used for user authentication:

```
SELECT * FROM users WHERE username = 'admin' AND password = 'password'
```

An attacker could inject a tautology into the password field, such as **' OR '1'='1'**. This would cause the query to become:

```
SELECT * FROM users WHERE username = 'admin' AND password = '' OR '1'='1'
```

Since the condition '1'='1' is always true, the query would return all users in the database, including the admin account. The attacker could then log in as the admin without knowing the correct password.

1.4.2 Error-based SQL injection

In error-based SQL injection, attackers can use SQL queries that trigger errors in order to learn about the structure and contents of a database. For example, an attacker could send an SQL query to the application that intentionally causes an error, and the error message returned by the database will contain information about the structure of the database or the content of the queried table.

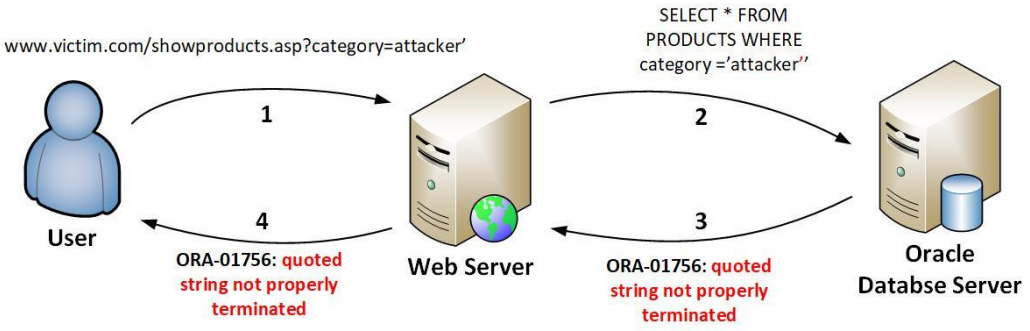

Figure 1.3 How Information flows during an SQL injection error [4].

According to Figure 1.3, the following occurs during a SQL injection error:

1. The user sends a request in an attempt to identify a SQL injection vulnerability. In this case, the user sends a value with a single quote appended to it.
2. The Web server retrieves user data and sends a SQL query to the database server. In this example, we can see that the SQL statement created by the Web server includes the user input and forms a syntactically incorrect query due to the two terminating quotes.
3. The database server receives the malformed SQL query and returns an error to the Web server.
4. The Web server receives the error from the database and sends an HTML response to the user. In this case, it sent the error message, but it is entirely up to the application how it presents any errors in the contents of the HTML response [4].

The following error is usually an indication of a MySQL injection vulnerability:

Warning: mysql_fetch_array(): supplied argument is not a valid MySQL result resource in /var/www/victim.com/showproduct.php on line 8

The preceding example illustrates the scenario of a request from the user which triggers an error in the database. Depending on how the application is coded, the response returned in step 4 will be constructed and handled as a result of one of the following:

- The SQL error is displayed on the page and is visible to the user from the Web browser.
- The SQL error is hidden in the source of the Web page for debugging purposes.
- Redirection to another page is used when an error is detected.
- An HTTP error code 500 (Internal Server Error) or HTTP redirection code 302 is returned.
- The application handles the error properly and simply shows no results, perhaps displaying a generic error page.

Mysql database errors

MySQL can be executed in many architectures and Operating systems. An Apache Web server running PHP on a Linux operating system forms a common configuration, but we can find it in many other scenarios as well.

In the example of Oracle error shown in Figure 1.3, the attacker injected a single quote in a GET parameter and the PHP page sent the SQL statement to the database. The following figure shows a fragment of MYSQL PHP code that displays the vulnerability [4]:

```php
<?php

//Connect to the database
$conn = new mysqli($servername, $username, $password, $dbname);

//Error checking in case the database is not accessible
if ($conn->connect_error) {
    die("Connection failed: " . $conn->connect_error);
}

//We retrieve category value from the GET request
$category = $_GET["category"];

//Create and execute the SQL statement
$sql = " SELECT * from products where category = '" . $category . "'";
$result = $conn->query($sql);

//Loop on the results
while ($row = $result->fetch_array(MYSQLI_NUM)) {
    printf("ID: %s Name: %s", $row[0], $row[1]);
}
//Free result set
$result->free_result();
```

Figure 1.4 SQL injection vulnerability in PHP code.

The code shows that the value retrieved from the GET variable is used in the SQL statement without sanitization. If an attacker injects a value with a single quote, the resultant SQL statement will be:

SELECT * FROM products WHERE category='attacker''

The preceding SQL statement will fail and the $conn->query($sql) function will not return any value. Therefore, the *$result* variable will not be a valid MySQL result resource. In the following line of code, the *$result -> fetch_array(MYSQLI_NUM)* function will fail and PHP

will show the warning message that indicates to an attacker that the SQL statement could not be executed.

In the preceding example, the application does not disclose details regarding the SQL error, and therefore the attacker will need to devote more effort in determining the correct way to exploit the vulnerability.

PHP has a built-in attribute of the *mysqli object* called *connect_error*, which provides information about the errors returned from the MySQL database during the execution of an SQL statement. In Figure 1.5, the following PHP code displays errors caused during the execution of the SQL query:

```
<?php

//Connect to the database

$conn = new mysqli($servername, $username, $password, $dbname);

//Error checking in case the database is not accessible
if ($conn->connect_error) {
    die("Connection failed: " . $conn->connect_error); }

//We retrieve category value from the GET request
$category = $_GET[“category”];

//Create and execute the SQL statement
$sql = "SELECT * from products where category='".$category."'";
$result = $conn->query($sql);

// Check if query executed without error
if ($result === false) {
    die("Error executing query: " . $conn->error);
}

//Loop on the results
while ($row = $result -> fetch_array(MYSQLI_NUM))
{   printf("ID: %s Name: %s", $row[0], $row[1]);

}
//Free result set
$result -> free_result();

$conn -> close(); ?>
```

Figure 1.5 handle query error with mysqli library in PHP.

When an application runs the preceding, the code that catches database errors and the SQL query fails, the returned HTML document will include the error returned by the database. If an attacker modifies a string parameter by adding a single quote the server will return output similar to the following:

Error: You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near ’’’at line 1

The preceding output provides information regarding why the SQL query failed. If the injectable parameter is not a string and therefore is not enclosed between single quotes, the resultant output would be similar to this:

Error: Unknown column ‘attacker’ in ‘where clause’

The behavior in MySQL server is identical to Microsoft SQL Server; because the value is not enclosed between quotes, MySQL treats it as a column name. The SQL statement executed was along these lines:

SELECT * FROM products WHERE idproduct=attacker

MySQL cannot find a column name called attacker and therefore returns an error.

This is the code snippet from the PHP script shown earlier in charge of error handling:

```
// Check if query executed without error
if ($result === false) {
    die("Error executing query: " . $conn->error);
}
```

In this example, the error is caught and then displayed using the *die()* function. The PHP *die()* function prints a message and gracefully exits the current script. Other options are available for the programmer, such as redirecting to another page:

```
<?php
//If there is any error
//Error checking and redirection
if ($result === false) {
    header("Location:http://www.victim.com/error.php");
}
```

1.4.3 Blind SQL Injection

Blind SQL injection is a technique of SQL Injection attack that asks the database true or false questions and determines the answer based on the application's response. This attack is often used when the web application is configured to show generic error messages but has not mitigated the code that is vulnerable to SQL injection.

When an attacker exploits SQL injection, sometimes the web application displays error messages from the database complaining that the SQL Query's syntax is incorrect. Blind SQL injection is nearly identical to normal SQL Injection, the only difference being the way the data is retrieved from the database. When the database does not output data to the web page, an attacker is forced to steal data by asking the database a series of true or false questions. This makes exploiting the SQL Injection vulnerability more difficult, but not impossible [5].

An attacker may verify whether a sent request returned true or false in a few ways:

1.4.3.1 Content-based

Using a simple page, which displays an article with a given ID as the parameter, the attacker may perform a couple of simple tests to determine if the page is vulnerable to SQL Injection attacks.

Example URL: http://newspaper.com/items.php?id=2, this URL access sends the following query to the database:

SELECT title, description, body FROM items WHERE ID = 2

The attacker may then try to inject a query that returns *'false'*:

http://newspaper.com/items.php?id=2 and 1=2

Now the SQL query should look like this:

SELECT title, description, body FROM items WHERE ID = 2 and 1=2

If the web application is vulnerable to SQL Injection, then it probably will not return anything. To make sure, the attacker will inject a query that will return 'true':

http://newspaper.com/items.php?id=2 and 1=1

If the content of the page that returns 'true' is different than that of the page that returns 'false', then the attacker is able to distinguish when the executed query returns true or false.

Once this has been verified, the only limitations are privileges set up by the database administrator, different SQL syntax, and the attacker's imagination.

1.4.3.2 Time-based

This type of blind SQL injection relies on the database pausing for a specified amount of time, then returning the results, indicating successful SQL query execution. Using this method, an attacker enumerates each letter of the desired piece of data using the following logic:

- *If the first letter of the first database's name is an 'A', wait for 10 seconds.*
- *If the first letter of the first database's name is a 'B', wait for 20 seconds. etc.*

Using some time-taking operation e.g. *BENCHMARK()*, will delay server responses if the expression is True.

BENCHMARK(5000000,ENCODE('MSG','by 5 seconds')) will execute the ENCODE function 5000000 times.

Depending on the database server's performance and load, it should take just a moment to finish this operation. The important thing is, from the attacker's point of view, to specify a high-enough number of BENCHMARK() function repetitions to affect the database response time in a noticeable way. Here is an example combination of both queries:

```
1 UNION SELECT IF(SUBSTRING(user_password,1,1) =
CHAR(50),BENCHMARK(5000000,ENCODE('MSG','by 5 seconds')),null) FROM users
WHERE user_id = 1;
```

If the database response took a long time, we may expect that the first user password character with *user_id = 1* is character '2', *(CHAR(50) == '2')*

Using this method for the rest of the characters, it's possible to enumerate entire passwords stored in the database. This method works even when the attacker injects the SQL queries and the content of the vulnerable page doesn't change.

Obviously, in this example, the names of the tables and the number of columns were specified. However, it is possible to guess them or check with a trial and error method.

Conducting Blind SQL Injection attacks manually is very time-consuming, but there are a lot of tools that automate this process. One of them is "SQLMap" partly developed within the OWASP grant program.

If the attacker is able to determine when their query returns True or False, then they may fingerprint the DBMS. This will make the whole attack much easier. If the time-based approach is used, this helps determine what type of database is in use. Another popular method to do this is to call functions, which will return the current date. MySQL, MSSQL, and Oracle have different functions for that, respectively *now()*, *getdate()*, and *sysdate()* [5].

1.4.4 Union-based SQL injections

The UNION operator is used in SQL to combine the results of two or more SELECT statements into a single result set. When a web application contains a SQL injection vulnerability that occurs in a SELECT statement, the attacker can often employ the UNION operator to perform a second, entirely separate query, and combine its results with those of the first.

All major DBMS products support UNION. It is the quickest way to retrieve arbitrary information from the database in situations where query results are returned directly [6]. Example of an application that enabled users to search for books based on author, title, publisher, and other criteria. Searching for books published by Wiley causes the application to perform the following query:

```
SELECT author, title, year FROM books WHERE publisher = 'Wiley'
```

Suppose that this query returns the following set of results:

AUTHOR	TITLE	YEAR
Litchfield	The Database Hacker's Handbook	2005
Anley	The Shellcoder's Handbook	2007

Table 1.1 Results of a SELECT query without UNION.

A far more interesting attack would be to use the *UNION* operator to inject a second SELECT query and append its results to those of the first. This second query can extract data from a different database table.

For example, entering the search term:

Wiley' UNION SELECT username,password,uid FROM users-- causes the application to perform the following query:

SELECT author, title, year FROM books WHERE publisher = 'Wiley' UNION SELECT username ,password, uid FROM users--'

This returns the results of the original search followed by the contents of the users table:

AUTHOR	TITLE	YEAR
Litchfield	The Database Hacker's Handbook	2005
Anley	The Shellcoder's Handbook	2007
admin	r00tr0x	0
cliff	Reboot	1

Table 1.2 Result of user query after a Union based SQL injection.

This simple example demonstrates the potentially huge power of the UNION operator when employed in a SQL injection attack. However, before it can be exploited in this way, two important provisos need to be considered [6]:

- When the results of two queries are combined using the UNION operator, the two result sets must have the same structure. In other words, they must contain the same number of columns, which have the same or compatible data types, appearing in the same order.
- To inject a second query that will return interesting results, the attacker needs to know the name of the database table that he wants to target, and the names of its relevant columns.

1.5 SQL Injection defense techniques

With user input channels being the main vector for such attacks, the best approach is controlling and vetting user input to watch for attack patterns by applying the following main prevention methods [7]:

1.5.1 Escaping

The developer must use always character-escaping functions for user-supplied input provided by each database management system (DBMS). This is done to make sure the DBMS never confuses it with the SQL statement provided by the developer.

For example, use the mysqli_real_escape_string() in PHP to avoid characters that could lead to an unintended SQL command. A modified version for the login bypass scenario would look like the following in Figure 1.6:

```
<?php

$db_connection = new mysqli($servername, $username, $password, $dbname);

$username = $db_connection->real_escape_string($_POST['username']);

$password = $db_connection->real_escape_string($_POST['password']);

$query = "SELECT * FROM users WHERE username = '" . $username . "' AND
password = '" . $password . "'";
```

Figure 1.6 Character-escaping in PHP code example.

Previously, the code would be vulnerable to adding an escape character (\) in front of the single quotes. However, having this small alteration will protect against an illegitimate user and mitigate SQL injection.

1.5.2 Input validation

The validation process is aimed at verifying whether or not the type of input submitted by a user is allowed. Input validation makes sure it is the accepted type, length, format, and so on. Only the value that passes the validation can be processed. It helps counteract any commands inserted in the input string. In a way, it is similar to looking to see who is knocking before opening the door.

Validation should not only be applied to fields that allow users to type in input, meaning developers should also take care of the following situations in equal measure:

- Use regular expressions as whitelists for structured data (such as name, age, income, survey response, zip code) to ensure strong input validation.
- In case of a fixed set of values (such as drop-down list, radio button), determine which value is returned. The input data should match one of the offered options exactly.

The below shows how to carry out table name validation.

```
switch ($tableName) {

    case 'fooTable': return true;

    case 'barTable': return true;

    default: return new Exception('unexpected value provided as table
name');
```

Figure 1.7 Input validation of a String example in PHP code.

The *$tableName* variable can then be directly appended—it is now widely known to be one of the legal and expected values for a table name.

In the case of a drop-down list, it's very easy to validate the data. Assuming the developers want a user to choose a rating from 1 to 5, he will change the PHP code to something like this:

```
<?php
if (isset($_POST["selRating"])) {

    $number = $_POST["selRating"];
    if ((is_numeric($number)) && ($number > 0) && ($number < 6)) {

        echo "Selected rating: " . $number;
    } else {
        echo "The rating has to be a number between 1 and 5!";
    }
```

Figure 1.8 Input validation of an integer example in PHP code.

The developers have added two simple checks:

1. It has to be a number (the is_numeric() function).
2. He requires that $number to be bigger than 0 and smaller than 6, which leaves him with a range of 1–5.

Data that is received from external parties has to be validated. This rule applies not only to the input provided by Internet users but also to suppliers, partners, vendors, or regulators. These vendors could be under attack and send malformed data even without their knowledge.

1.5.3 Parameterized queries

Parameterized queries are a means of pre-compiling an SQL statement so that you can then supply the parameters in order for the statement to be executed. This method makes it possible for the database to recognize the code and distinguish it from input data.

The user input is automatically quoted and the supplied input will not cause a change of the intent, so this coding style helps mitigate an SQL injection attack.

PHP 5.1 up versions present a better approach when working with databases: PHP Data Objects (PDO). PDO adopts methods that simplify the use of parameterized queries. Additionally, it makes the code easier to read and more portable since it operates on several databases, not just MySQL. The code in Figure 1.9 uses PDO with parameterized queries to prevent the SQL injection vulnerability:

```
<?php
$id = $_GET['id'];
$db_connection =
    new PDO('mysql:host=localhost;dbname=sql_injection_example', 'dbuser',
'dbpasswd');

//preparing the query
$sql = "SELECT username FROM users WHERE id = :id";
$query = $db_connection->prepare($sql);
$query->bindParam(':id', $id);
$query->execute();

//getting the result
$query->setFetchMode(PDO::FETCH_ASSOC);
$result = $query->fetchColumn();
print(htmlentities($result));
```

Figure 1.9 Parameterized queries using PDO in PHP code example.

1.5.4 Web application firewalls WAF

One of the best practices to identify SQL injection attacks is having a web application firewall (WAF). A WAF operating in front of the web servers monitors the traffic which goes in and out of the web servers and identifies patterns that constitute a threat. Essentially, it is a barrier put between the web application and the Internet.

A WAF operates via defined customizable web security rules. These sets of policies inform the WAF what weaknesses and traffic behavior it should search for. So, based on that information, a WAF will keep monitoring the applications and the GET and POST requests it receives to find and block malicious traffic [7].

WAFs provide efficient protection from a number of malicious security attacks such as:

- SQL injection
- Cross-site scripting (XSS)
- Session hijacking
- Distributed denial of service (DDoS) attacks
- Cookie poisoning
- Parameter tampering

1.5.5 Detection using machine learning

Artificial **I**ntelligence can enhance the speed and efficiency of SQL injection detection. By automating the process of query analysis and classification, AI can rapidly scan vast amounts of incoming queries, flagging potential threats in real-time.

SQL injections can be detected using machine-learning algorithms that are commonly used for binary classification tasks; they aims to find an optimal hyperplane that separates the legitimate queries from the malicious ones by maximizing the margin between them.

To train a model using machine learning, a systematic process that include **Dataset Preparation**, **Feature Extraction**, **Dataset Splitting**, **and Model Training** must be followed. During the training point, the model learns the decision boundary based on the provided features and their corresponding labels.

1.6 Conclusion

As web applications become increasingly complex, SQL injection attacks remain a persistent and evolving threat. However, various techniques and tools have been developed to detect and prevent SQL injection attacks like input validation and sanitization, parameterized queries, and security-focused coding practices.

Unfortunately, because of the large variation in the pattern of SQL injection attacks, it is often unable to protect databases. Therefore, it is recommended, to apply the above-mentioned techniques in combination with Machine Learning and AI tools.

Chapter 2

Deep Learning

2.1 Introduction

Artificial intelligence (AI) is a field of computer science that aims to create intelligent systems capable of performing tasks that typically require human intelligence. From self-driving cars to voice assistants, AI has made remarkable strides, transforming the way we live, work, and interact with technology.

In this chapter, we will explore common methods of machine learning, including supervised, unsupervised, and reinforcement learning. We will also discuss popular algorithmic approaches in machine learning, moving to explore deep learning and its architectures. Our focus will be on understanding the fundamental concepts behind these methods and algorithms.

2.2 Machine learning

The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience. In recent years many successful machine learning applications have been developed, ranging from data-mining programs that learn to detect fraudulent credit card transactions to information-filtering systems that learn users' reading preferences, to autonomous vehicles that learn to drive on public highways. At the same time, there have been important advances in the theory and algorithms that form the foundations of this field [8].

2.2.1 Types of machine learning

Machine learning can be categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning.

2.2.1.1 Supervised learning

Supervised learning involves training machine learning models on labeled data. The model learns to make predictions or classify new data by finding patterns and relationships between the input features and the desired outputs. Examples of supervised learning algorithms include linear regression, decision trees, and support vector machines (SVMs).

Supervised learning addresses two types of problems: classification problems and regression problems.

Classification is a type of supervised learning problem where the goal is to predict a discrete class label for each data point. For example, the goal might be to predict whether an image contains a cat or a dog, or whether a patient has cancer or not. Classification problems can be solved using a variety of machine learning algorithms, including decision trees and support vector machines.

Regression is a type of supervised learning problem where the goal is to predict a continuous value for each data point. For example, the goal might be to predict the price of a house or the height of a person. Regression problems can be solved using a variety of machine learning algorithms, including linear regression and polynomial regression.

2.2.1.2 Unsupervised learning

Unsupervised learning involves training models on unlabeled data. The algorithms learn to identify patterns, structures, and relationships in the data without explicit guidance. Common tasks in unsupervised learning include clustering, where similar data points are grouped together, and dimensionality reduction, which aims to reduce the complexity of data while preserving important information. Examples of unsupervised learning algorithms include k-means clustering, hierarchical clustering, and principal component analysis (PCA) [10].

2.2.1.3 Reinforcement

Reinforcement learning involves training agents to interact with an environment and learn optimal behaviors through trial and error. The agent receives feedback in the form of rewards or penalties based on its actions, and its objective is to maximize cumulative rewards over time. Reinforcement learning is commonly used in scenarios where an agent needs to make sequential decisions, such as in game playing, robotics, and autonomous vehicle control. Popular

reinforcement learning algorithms include Q-learning, policy gradients, and deep Q-networks (DQNs).

It's worth noting that these types of machine learning are not mutually exclusive, and hybrid approaches that combine multiple types can be used to tackle complex problems. Additionally, there are other specialized areas within machine learning, such as semi-supervised learning and transfer learning, which further expand the capabilities of machine learning algorithms [11].

2.2.2 Machine learning algorithms

Machine learning algorithms can be thought of as powerful tools that allow us to unlock the potential of data. By applying these algorithms to various problems, we can uncover hidden patterns, detect anomalies, classify data into different categories, and even make accurate predictions about future outcomes.

2.2.2.1 Linear Regression

Linear regression is a machine learning algorithm that uses a line to predict a numerical value based on input variables. As shown in Figure 2.1 the line is found by minimizing the difference between the predicted and actual values. Linear regression is simple to understand and interpret, and it is commonly used in a variety of domains [9].

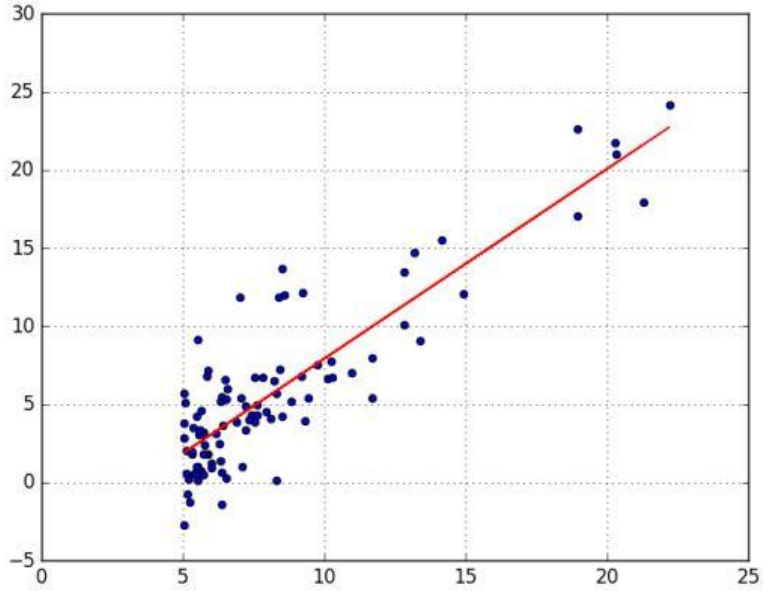

Figure 2.1 Graphical representation of linear regression.

2.2.2.2 Logistic Regression

Logistic regression is a machine-learning algorithm that predicts the probability of an instance belonging to a specific class. It is a supervised learning method that uses the logistic function to model the relationship between the input variables and the probability of the binary outcome as shown in Figure 2.2. Logistic regression is commonly used in a variety of domains, including customer behavior analysis, fraud detection, and medical diagnosis. Its advantages lie in its simplicity, interpretability, and ability to handle linear and nonlinear relationships between variables [9].

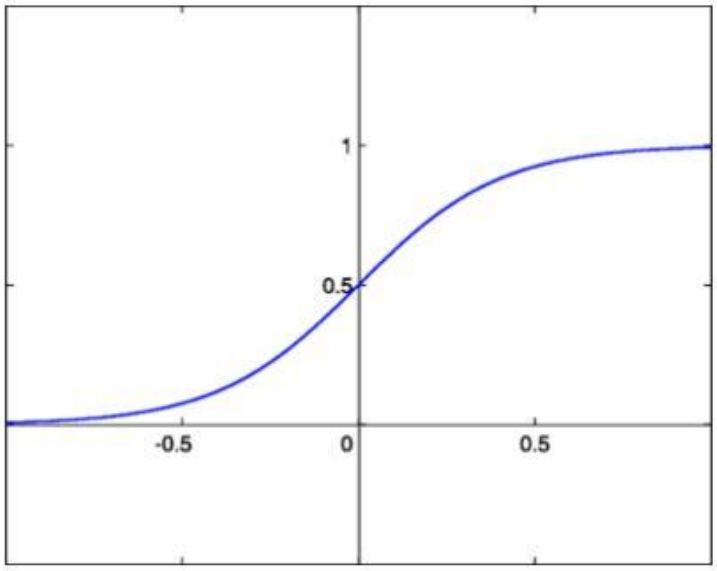

Figure 2.2 Graphical representation of logistic regression.

2.2.2.3 Support vector machines

Support vector machines (SVMs) are a machine learning algorithm that can be used for both classification and regression tasks. As shown in Figure 2.3 SVMs work by finding the optimal hyperplane that separates data points of different classes with the maximum margin. This means that the hyperplane is as far away as possible from any data points on either side. SVMs can handle both linearly separable and non-linearly separable data by using kernel functions to implicitly map the data into a higher-dimensional space.

Their key strengths lie in their ability to handle high-dimensional data and flexibility in choosing different kernel functions [9].

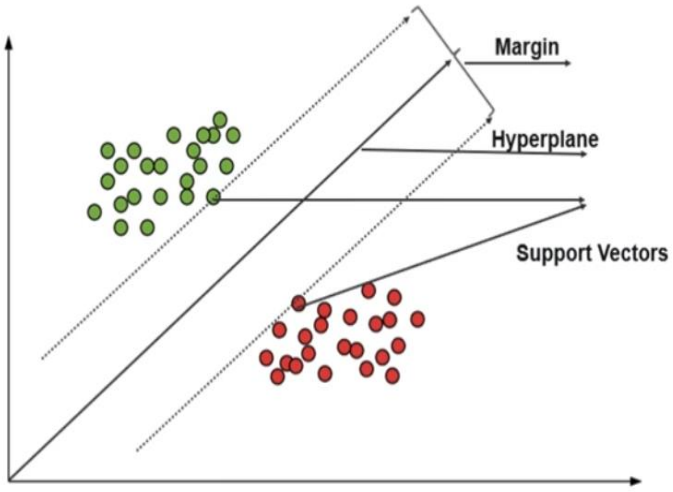

Figure 2.3 Graphical representation of Support vector machines.

2.2.2.4 K-Means

K-means is an unsupervised machine learning algorithm that groups data points into clusters based on their similarities. The algorithm works by iteratively assigning data points to the cluster with the nearest centroid, and then updating the centroids based on the mean of the assigned points as shown in Figure 2.4. The number of clusters, K, is predefined by the user. K-means is a simple and computationally efficient algorithm, making it widely used for clustering tasks.

K-means finds applications in various domains, including customer segmentation, image compression, and document clustering.

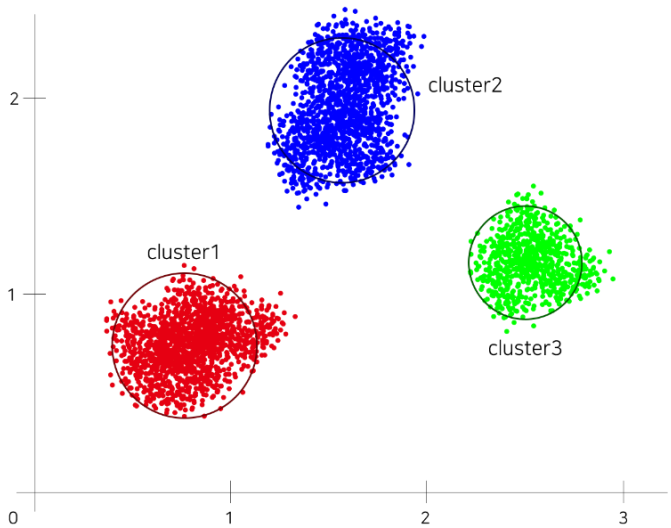

Figure 2.4 Graphical representation of k means.

2.2.3 Machine learning applications

Machine learning is a rapidly evolving field with a wide range of applications. It has been used to improve healthcare, transportation, finance, retail, energy, and agriculture.

Healthcare and Medicine: In healthcare, machine learning is used to diagnose diseases, recommend treatments, and personalize care.

Automotive Industry: In transportation, machine learning is used to develop autonomous vehicles and improve traffic management.

Financial Services: In finance, machine learning is used to detect fraud, assess risk, and make investment decisions.

Energy and Utilities: In energy, machine learning is used to optimize energy consumption, predict outages, and improve grid reliability.

Agriculture: In agriculture, machine learning is used to predict crop yields, detect pests, and optimize irrigation.

These are just some of the industries that machine learning had been applied to, machine learning is a powerful tool that has the potential to transform many industries.

2.3 Deep learning

Deep learning (DL) is a subset of machine learning (ML) that uses multilayer artificial neural networks to model and solve complex problems. These networks are designed to simulate the behavior of neurons in the human brain, enabling them to process and learn from large amounts of unstructured data.

Deep learning algorithms automatically learn to recognize patterns and features in data by analyzing and adjusting the weights and biases of network interconnected nodes. This process, called training, involves optimizing network parameters to minimize errors and improve accuracy.

Deep learning has grown in popularity in recent years due to its ability to process large and diverse datasets, achieve cutting-edge performance in many fields, and achieve breakthroughs in areas such as computer vision, natural language processing, and speech recognition [12].

2.3.1 Artificial neural networks

Typical artificial neural networks (ANN) are biologically inspired computer programs inspired by the workings of the human brain. These ANNs are referred to as networks because they are made up of several functions, which collect knowledge by recognizing links and patterns in data using previous experiences referred to as training examples in most literature. The learned patterns in data are adjusted by an appropriate activation function and displayed as the neuron's output.

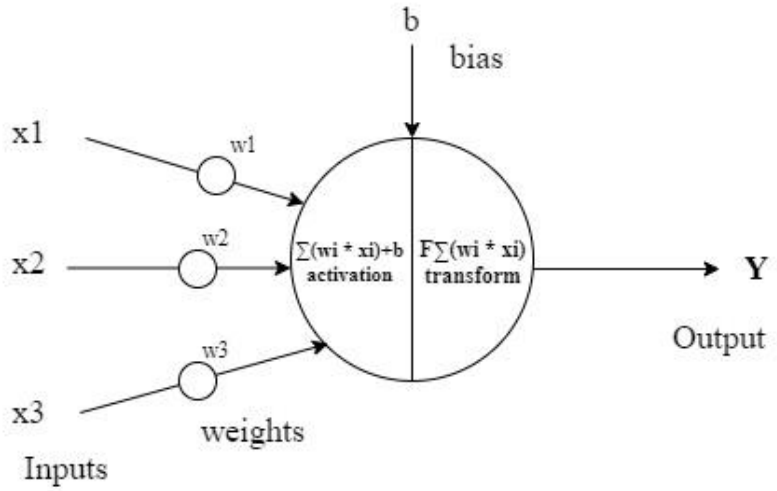

Figure 2.5 Typical biological-inspired neuron.

The Figure 2.5 represents a typical biological-inspired neuron with three inputs, denoted as x1, x2, and x3. These inputs are connected to an activation function represented by a circle, which plays a crucial role in determining the output of the neural network. The activation function takes the weighted sum of the inputs and applies a non-linear transformation to produce the final output.

The weights w1, w2, and w3 represent the respective strengths or importance assigned to each input. These weights are multiplied with their corresponding inputs and then summed up. The purpose of the weights is to control the impact of each input on the output of the network. By adjusting the values of these weights during the learning process, the neural network can learn to make accurate predictions or classifications based on the given inputs. a bias is an additional parameter that is added to the weighted sum of inputs before passing through the activation function. Bias allows the neural network to make adjustments to the output even when all the input values are zero.

The activation function serves as a decision-making element. It takes the weighted sum of the inputs and applies a specific mathematical function to determine the output of the neural network.

Finally, the output of the activation function represents the result or prediction of the artificial neural network. It can be used for various tasks, such as classification, regression, or pattern recognition, depending on the problem being addressed. The goal of training the neural network is to adjust the weights and biases in such a way that it produces accurate outputs for a given set of inputs.

ANNs are generally composed of many layers, including an input layer, one or more hidden layers, and an output layer. ANNs can catch complicated patterns and generate accurate predictions because of their layer-wise structuring.

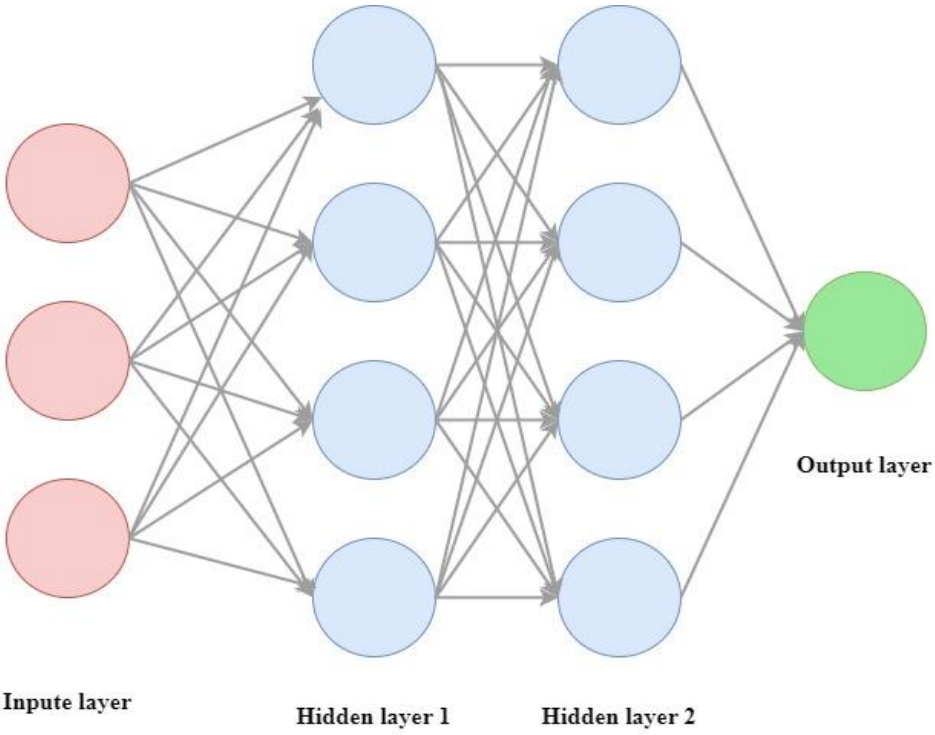

Figure 2.6 Schematic representation of a neural network.

The **input layer** serves as the entry point for the neural network, receiving the initial input values. Each input node represents a feature or attribute of the problem being solved. For example, in an image recognition task, each input node could represent a pixel value. The input layer simply transmits these values to the next layer.

Hidden layers, as the name suggests, are the intermediate layers between the input and output layers, where the processing and transformation of information occur. These layers are called "hidden" because their computations are not directly observable from the outside. Each hidden layer consists of multiple artificial neurons (also called nodes or units) that receive inputs from the previous layer, perform calculations, and pass the results to the next layer. The number of hidden layers and the number of neurons in each hidden layer can vary depending on the complexity of the problem and the desired network architecture. Additional hidden layers and neurons allow the network to learn more intricate representations and potentially improve its performance. Deep neural networks, which have multiple hidden layers, have been successful in solving tasks such as image recognition, natural language processing, and speech recognition.

Finally, we have the **output layer**, which produces the final result or prediction of the neural network. The number of nodes in the output layer corresponds to the number of possible outputs or classes in the problem. For instance, in a binary classification task, there would be two output nodes representing the two possible classes. In a regression task, the output layer might consist of a single node that produces a continuous value.

The connections between the layers, represented by **weights**, determine the strength and influence of information flowing through the network. During the training phase, these weights are adjusted through a process called backpropagation, which involves propagating the error from the output layer back to the hidden layers and adjusting the weights accordingly. This iterative learning process allows the neural network to gradually optimize its performance and improve its ability to make accurate predictions.

2.3.2 Activation functions

Activation functions are an essential component of neural networks as they introduce non-linearity, allowing the network to learn and approximate complex patterns in the data. Here are some commonly used activation functions:

Sigmoid: The sigmoid function is a smooth S-shaped curve that maps the input to a value between 0 and 1. It is often used in binary classification problems or as an output activation function in models that require probability-like outputs. Mathematically sigmoid is represented as:

$$f(t) = \frac{1}{(1 + e^{-t})}$$

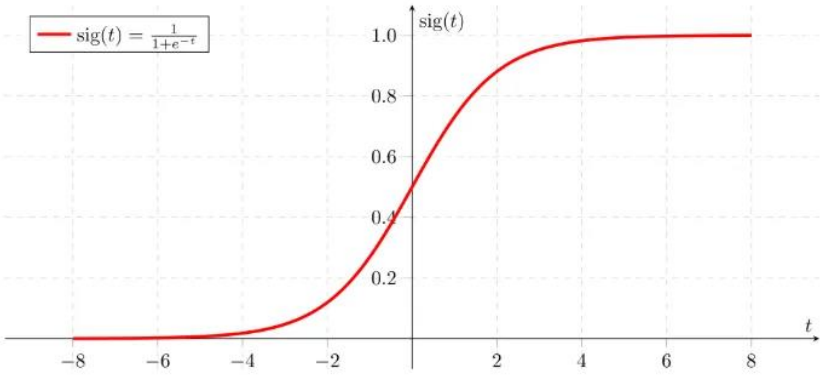

Figure 2.7 Sigmoid activation function.

Softmax: The softmax is a more generalized form of the sigmoid, it is a specialized activation function used in multi-class classification problems. It takes a vector of real numbers as input and normalizes them into a probability distribution over multiple classes, where the sum of the probabilities is 1. for each element zi of the input vector z, mathematically Softmax is represented as:

$$\mathrm{f(x)} = \frac{\exp\ (xi)}{\Sigma \exp\ (\mathrm{xj})}$$

ReLU (Rectified Linear Unit): ReLU is a piecewise linear function that returns the input as if it is positive, and 0 otherwise. It is widely used in hidden layers of deep neural networks due to its simplicity and computational efficiency. Mathematically ReLU is represented as:

$$f(x) = max(0, x)$$

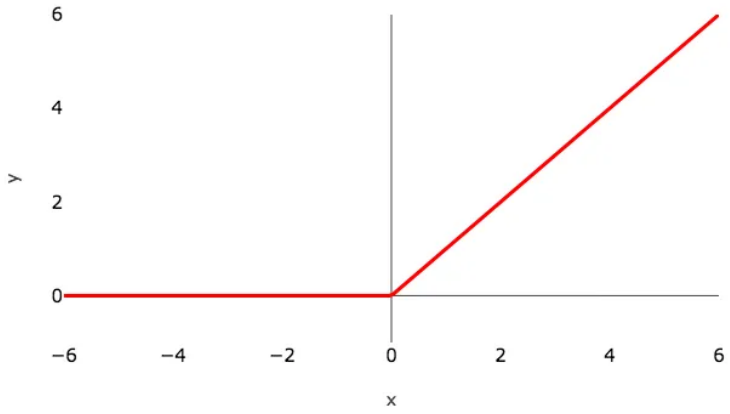

Figure 2.8 ReLU activation function.

Tanh (Hyperbolic Tangent): The hyperbolic tangent function is similar to the sigmoid function but maps the input to a value between -1 and 1. It provides stronger non-linearity than the sigmoid function and is often used in recurrent neural networks (RNNs) and convolutional neural networks (CNNs). Mathematically Tanh is represented as:

$$f(x) = \frac{(e^{x} - e^{-x})}{(e^{x} + e^{-x})}$$

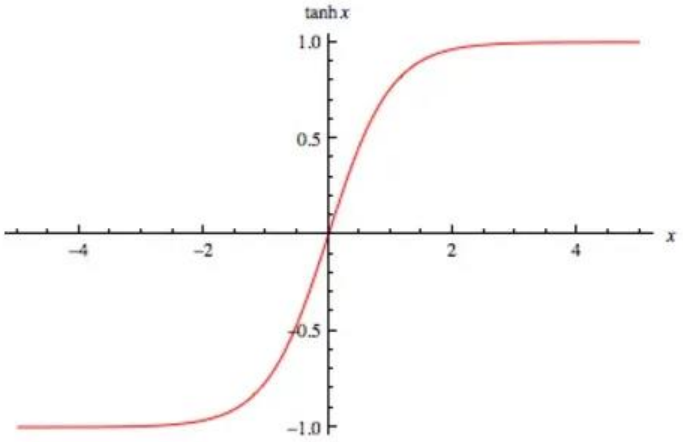

Figure 2.9 Tanh activation function.

These activation functions serve different purposes and may be more suitable for specific tasks or network architectures. Choosing the appropriate activation function depends on the nature of the problem and the desired behavior of the neural network.

2.3.3 Deep learning architectures

Deep Learning is a growing field with applications that span across several use cases. In each use case, a different architecture is predominant and gives the best efficiency.

There are many different types of deep learning architectures, many of which are derived from original architectures. Some of the most popular ones are Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory Networks (LSTMs).

For the purposes of this discussion, we will talk about Recurrent Neural Networks, Long Short-Term Memory Networks, Gated Recurrent Units (GRU) and focus on one type of

architecture known as transformers, which have gained popularity in recent years for their ability to process sequential data with parallelization and attention mechanisms.

2.3.3.1 Recurrent Neural Networks

Recurrent neural networks (RNNs) are a type of neural network that is particularly effective in processing sequential data. Unlike traditional neural networks, which process input data independently, RNNs have a feedback mechanism that allows them to retain and utilize information from previous steps in the sequence. This makes them well-suited for tasks such as natural language processing, speech recognition and machine translation [12].

RNNs work by processing each step in the sequence one at a time. At each step, the RNN takes as input the current input data and the output from the previous step. The RNN then uses this information to calculate a new output. This output is then used as input for the next step, and so on [14].

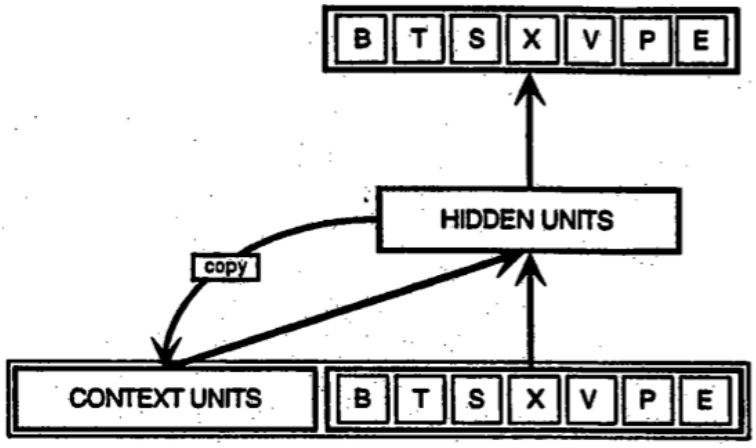

Figure 2.10 Diagram of simple recurrent network.

The diagram shown in Figure 2.10 of an early, simple recurrent network, where the BTSXVPE at the bottom of the drawing represents the input example in the current moment, and CONTEXT UNIT represents the output of the previous moment.

Some popular variants of RNNs include LSTMs (Long Short-Term Memory) and GRUs (Gated Recurrent Units). LSTMs and GRUs are designed to address the vanishing gradient problem that can occur in traditional RNNs, which hinders their ability to capture long-term dependencies. LSTMs incorporate memory cells and gating mechanisms that enable them to selectively retain and update information over time. GRUs have gating mechanisms that control the flow of information within the network, allowing them to capture long-term dependencies efficiently [15].

2.3.3.2 Long Short-Term Memory Networks

Traditional RNNs suffer from a problem known as the vanishing gradient problem, which can hinder their performance when dealing with long-term dependencies.

The vanishing gradient problem arises during the training of RNNs when the gradients used for learning become extremely small as they propagate backward through time. This occurs because the gradient calculation involves multiplying a series of weight matrices, and the gradient values tend to diminish exponentially with each multiplication. As a result, the network struggles to capture and propagate information from earlier time steps, limiting its ability to model long-term dependencies in the data [16].

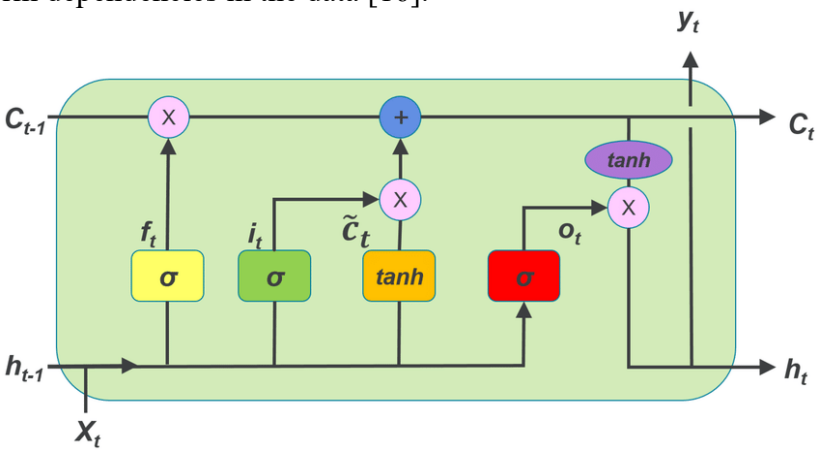

Figure 2.11 Long Short-term Memory Neural Network.

To address the vanishing gradient problem, the Long Short-Term Memory (LSTM) architecture was introduced. LSTM networks incorporate specialized memory cells that allow for better information retention and controlled information flow.

The key innovation of LSTM is the inclusion of memory cells, which serve as a way to store and access information over long time scales. These memory cells are equipped with three main components: an input gate, a forget gate, and an output gate. These gates regulate the flow of information into and out of the memory cells, allowing them to selectively retain and forget information as needed.

The input gate determines how much new information is stored in the memory cells, while the forget gate controls which information is discarded. The output gate determines the information to be passed on to the next time step. By carefully controlling the flow of information, LSTM networks can effectively address the vanishing gradient problem and capture long-term dependencies in sequential data [17].

The LSTM architecture has been widely successful in various applications, including natural language processing, speech recognition, machine translation, and video analysis. Its ability to model long-term dependencies makes it a crucial component in many state-of-the-art deep learning models.

2.3.3.3 Gated Recurrent Units

The Gated Recurrent Unit (GRU) is a type of recurrent neural network (RNN) architecture that addresses the vanishing gradient problem of traditional RNNs. GRUs were introduced in 2014 by Cho et al. to capture long-term dependencies in sequential data while mitigating the issue of information loss over time.

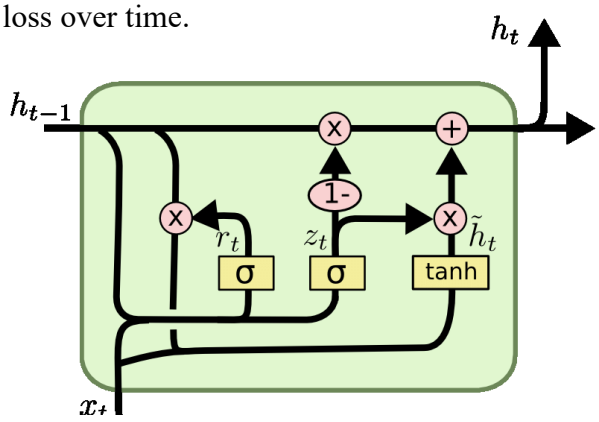

Figure 2.12 Gated Recurrent Unit.

GRUs consist of two main gates: the reset gate and the update gate. The reset gate determines how much of the previous hidden state should be forgotten, while the update gate determines how much of the new candidate state should be incorporated into the current hidden state. These gates allow GRUs to selectively retain and update information over time, facilitating the learning of complex dependencies in sequential data.

GRUs have several advantages over traditional RNNs and even other gated architectures like LSTMs. They have fewer parameters, making them computationally more efficient and easier to train. Additionally, GRUs have been found to perform competitively or even outperform LSTMs on certain tasks, while requiring less training time and data [18].

Overall, GRUs are a powerful and versatile RNN architecture that offers several advantages over traditional RNNs and LSTMs.

2.3.3.4 Transformers

As we already mentioned that RNNs have several limitations, including difficulty in parallelization and difficulty in capturing long-term dependencies. These limitations have led to the development of alternative models, such as transformers.

The Transformer is a neural network architecture that was proposed by Vaswani in 2017. It is a powerful model that has revolutionized the field of natural language processing (NLP). Transformer models introduced a new approach to sequence modeling without recurrent connections, which makes them more efficient and easier to train than traditional RNN-based models. Transformer models have been widely adopted and have become the state-of-the-art models in various NLP tasks, including machine translation, question answering, and text generation [19].

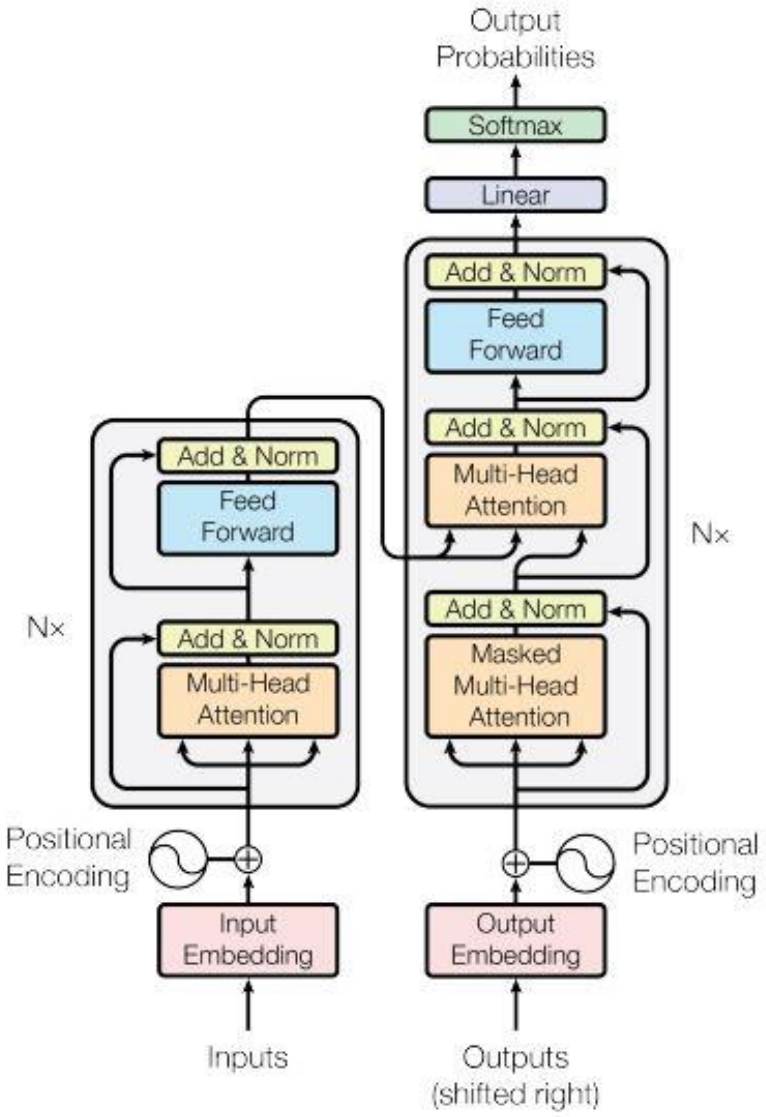

Figure 2.13 Architecture of transformers [19].

The Transformer architecture adopts an encoder-decoder structure. As shown in Figure 2.13, the encoder maps an input sequence of symbol representations (x1, ..., xn) to a sequence of continuous representations (z1, ..., zn). The decoder, using the encoded representation z, generates an output sequence (y1, ..., ym) by producing one symbol at a time. This process is auto-regressive, as the model incorporates previously generated symbols as additional input for generating the next symbol. The Transformer architecture follows this overall design, employing stacked self-attention and point-wise, fully connected layers for both the encoder and decoder [19].

Encoder

The encoder is made up of a stack of N = 6 identical layers. Each layer has two sub-layers. The first sub-layer is a multi-head self-attention mechanism, and the second is a simple feed-forward network. Each of these two layers is followed by a normalization layer. To facilitate the connections in the model, all sub-layers produce outputs of dimention dmodel = 512.

Decoder

The decoder is also made up of a stack of N = 6 identical layers. Each layer has three sub-layers. In addition to the two sub-layers in each encoder the decoder inserts a third sub-layer which is a multi-head attention over the output of the encoder. Similar to the encoder each of the three sub-layers is followed by a normalization layer. In the decoder stack, the self-attention sub-layer is modified to make sure positions don't pay attention to the positions that come after them, using a technique called masking. This, along with the fact that the output embeddings are shifted by one position, ensures that the predictions for a particular position only rely on the already known outputs at positions before it.

Attention

An attention function can be defined as a mapping between a query and a set of key-value pairs, producing an output. In this mapping, the query, keys, values, and output are all vectors. The output is computed by calculating a weighted sum of the values, where the weight assigned to each value is determined by a compatibility function between the query and the corresponding key.

Scaled Dot-Product Attention

To compute the scaled dot-product attention, the query and key vectors are multiplied together to obtain a similarity score for each pair. The dot product represents how well the query aligns with each key. These scores are then scaled by the square root of the dimension of the key vectors to avoid overly large values.

Next, the scaled scores are passed through a softmax function, which normalizes the scores and assigns a weight to each value vector. The softmax function ensures that the weights sum up to 1, making the attention mechanism a proper distribution.

Finally, the values are multiplied by their corresponding weights and summed up to produce the output of the attention mechanism. This output captures the relevant information from the value vectors based on the similarity between the query and key vectors.

Multi-Head Attention

Instead of performing a single attention function with keys, values, and queries of dmodel dimensions, it has been found beneficial to linearly project the queries, keys, and values h times. Each projection uses different learned linear transformations to convert them into dq, dk, and dv dimensions, respectively. Subsequently, the attention function is applied in parallel to these projected versions of queries, keys, and values, resulting in dv-dimensional output values. These values are then concatenated and projected again, producing the final values.

Feed-Forward Networks

In addition to the attention sub-layers, each layer in the encoder and decoder includes a fully connected feed-forward network. This network operates on each position independently and uniformly. It comprises two linear transformations with a ReLU activation function in between [13].

Embeddings and Softmax

Similar to other sequence transduction models, learned embeddings are used to convert input tokens and output tokens into dmodel-dimensional vectors. The decoder output is transformed into predicted probabilities for the next token using a learned linear transformation and softmax function.

Positional Encoding

Positional encoding adds position-related information to the input data, allowing the model to understand the order of tokens. It involves incorporating fixed-length vectors into the input embeddings to capture sequential dependencies. This enables transformers to effectively process sequential data for tasks like machine translation and language understanding.

2.3.4 Deep learning applications

Deep learning (DL) is a powerful tool that has revolutionized many fields by delivering unprecedented accuracy and efficiency in processing complex data. One of the main strengths of DL is its ability to automatically extract relevant features and patterns from large and diverse datasets without the need for manual feature extraction. This makes deep learning particularly suitable for applications such as image and video processing, natural language understanding, speech recognition, and autonomous decision-making. Some of deep learning applications are:

Image recognition: Deep learning models can be used to identify objects in images. This technology is used in a variety of applications, such as facial recognition, self-driving cars, and medical image analysis.

Natural language processing: Deep learning models can be used to understand and process human language. This technology is used in a variety of applications, such as speech recognition, machine translation, and chatbots.

Machine translation: Deep learning models can be used to translate text from one language to another. This technology is used in a variety of applications, such as online translation services and multilingual software.

Robotics: Deep learning models can be used to control robots. This technology is used in a variety of applications, such as self-driving cars and industrial robots.

2.4 Conclusion

In conclusion, this chapter has provided an overview of both machine learning and deep learning. We have discussed their fundamental concepts, algorithms, and architectures, shedding light on these powerful techniques in the field of artificial intelligence.

Chapter 3

Conception and Implementation

3.1 Introduction

In this chapter, we will discuss the design and implementation of our model for detecting SQL injections using deep learning. We will describe the general conception of our work and the materials used, including the dataset and the type of deep learning architecture employed. Furthermore, we will cover the preprocessing steps taken to ensure the accuracy and efficiency of our system. By detailing our approach to design and implementation, we aim to provide a comprehensive understanding of our methodology for detecting SQL injections through the use of deep learning.

3.2 General conception of the solution

The deep learning model that we have developed will be used for detecting SQL injection attacks in web applications. The model is based on the Bidirectional Encoder Representations from Transformers (BERT) architecture, which has been fine-tuned on a dataset of SQL injection attacks and normal SQL queries. The model is designed to function as a middleware layer (API or a WAF) between the web application and the database server as shown in Figure 3.1

Our **Sqli Detection Model** analyzes incoming queries and detects any suspicious patterns that may indicate an SQL injection attack, it can be used on Web Application Firewalls (WAF) or as an API. Once the SQL injection is detected, the tool can either block the request or alert the system administrator, depending on the configuration.

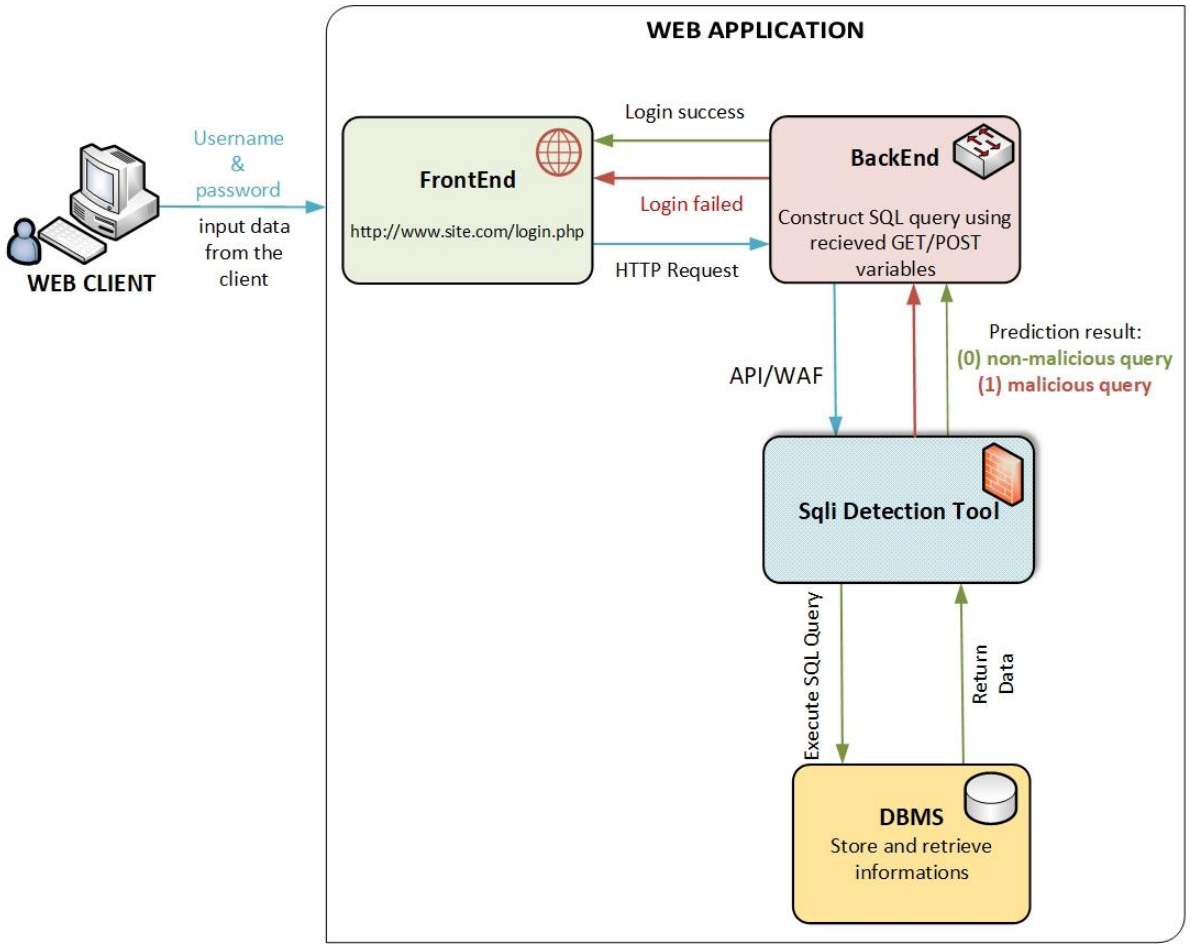

Figure 3.1 Sql injection Detection Tool conception and architecture.

3.3 Chosen model: BERT

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art language model that has revolutionized natural language processing tasks. It is based on transformer architecture, which enables it to effectively capture contextual information from input text. Unlike traditional models that process text sequentially, BERT takes into account the entire context surrounding each word by using a bidirectional approach. By pre-training on large amounts of unlabeled text data, BERT learns to generate high-quality word representations that encode rich semantic and syntactic information. These pre-trained representations can then be fine-tuned on specific tasks, such as detecting SQL injections in our case. With its ability to understand the nuanced context of language, BERT has demonstrated

exceptional performance on various natural language processing tasks, making it a suitable choice for enhancing the detection and prevention of SQL injection attacks in this study.

BERT has two variants: BERT-base and BERT-large, which differ in the number of layers and parameters. BERT-base has 12 transformer layers and 110 million parameters, while BERT-large has 24 transformer layers and 340 million parameters [20].

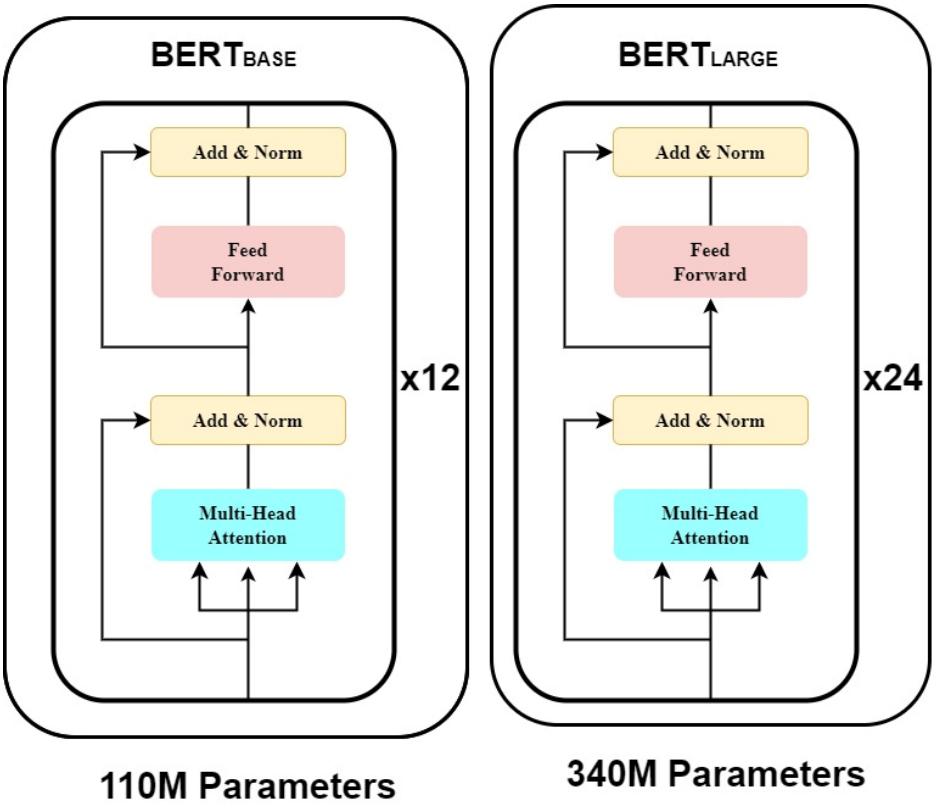

Figure 3.2 BERT model size.

In this study, the choice of the BERT-Base model was driven by the size of our dataset, which consisted of approximately 22,000 samples. Given the limited size of the dataset, using BERT-Large would not have been the most suitable option.

3.3.1 BERT architecture

BERT uses the Transformer architecture, which is an attention mechanism designed to learn contextual relationships between words or sub-words in a text. In its original form, the Transformer consists of two mechanisms: an encoder that processes the text input and a decoder that generates predictions for the task. However, since BERT's objective is to generate a language model, only the encoder mechanism is necessary.

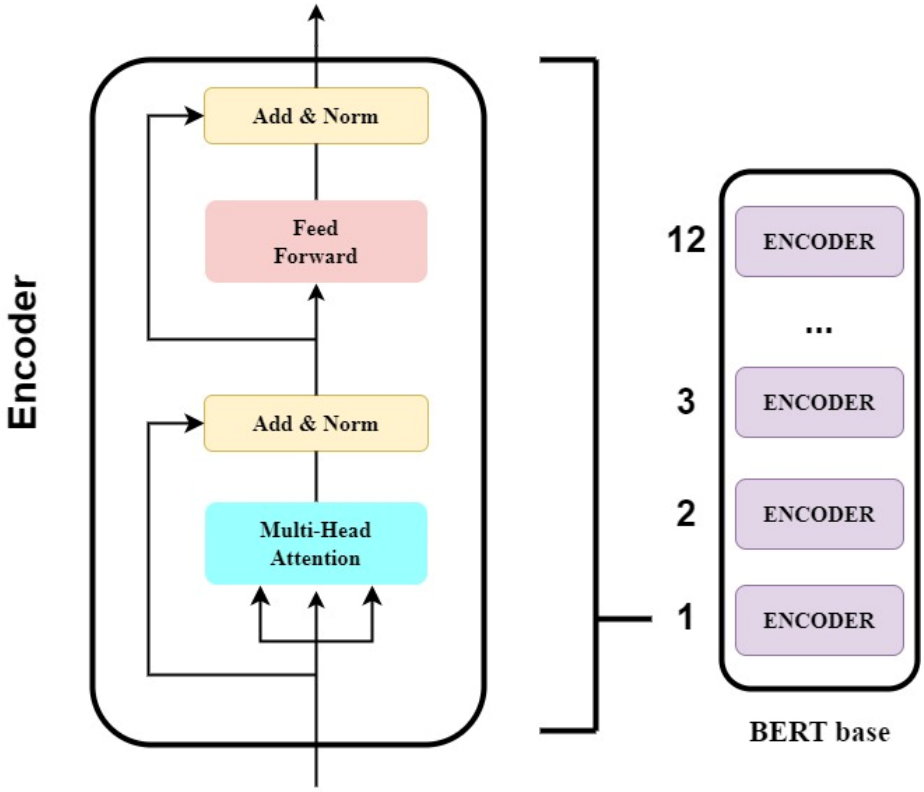

Figure 3.3 BERT model architecture.

The encoder in BERT comprises a stack of N = 6 identical layers, with each layer consisting of two sub-layers. The first sub-layer is a multi-head self-attention mechanism, which allows the model to attend to different positions within the input sequence simultaneously. This mechanism employs multiple attention heads to capture dependencies between words or sub-words in a text. The second sub-layer is a simple feed-forward network, which processes the outputs from the previous self-attention layer. This network consists of fully connected layers and applies a linear transformation to each position independently. Both sub-layers in each encoder layer are followed by a normalization layer. This normalization helps in stabilizing the learning process by normalizing the outputs of each sub-layer. Additionally, to ensure smooth connectivity between layers in the model, all sub-layers produce outputs of dimension dmodel = 512, ensuring consistent input and output dimensions throughout the architecture.

3.3.2 BERT for Text Classification

When processing input, BERT expects a sequence of tokens with a maximum length of 512 tokens. The sequence can be divided into one or two segments. The first token of the sequence

is a special token that holds a special classification embedding. For text classification tasks, BERT considers the final hidden state h of the first token as the representation of the entire sequence. This hidden state captures the information from the entire input sequence and serves as a condensed representation. To make predictions, a simple softmax classifier is added on top of BERT. The classifier employs a task-specific parameter matrix W to compute the probability of each label c given the representation h.

3.3.3 Why BERT was chosen

In the context of SQL injection detection, BERT was chosen for its ability to learn rich representations of text data, including queries, comments, and other textual elements that are typically associated with SQL injection attacks. BERT has been shown to outperform other state-of-the-art models on various NLP tasks [21], making it a promising candidate for SQL injection detection. Its transformer-based architecture allows it to process input sequences in a bidirectional manner and generate contextualized word representations. Fine-tuning BERT on a labeled dataset of SQL queries allowed us to develop a model that can detect SQL injection attacks with high accuracy.

3.3.4 Fine-tuning BERT for SQL Injection Detection:

To use BERT for SQL injection detection, we fine-tuned the pre-trained BERT model on a labeled dataset of normal SQL queries and SQL injections. During fine-tuning, the model was trained to predict whether a given query is a SQL injection or not. The fine-tuning process involved adjusting the weights of the classification layer while keeping the weights of the pre-trained BERT layers fixed.

3.4 Presentation of development tools

3.4.1 Programming language

3.4.1.1 Python 

Python is a popular programming language in the field of machine learning and artificial intelligence due to its simple syntax, extensive libraries, and ease of use. Python provides a variety of libraries for machine learning such as TensorFlow, PyTorch, and Keras, making it a go-to language for many data scientists and machine learning practitioners. Its libraries provide an extensive range of functionalities, from data preprocessing to complex neural network architectures [22].

Furthermore, Python's community is continuously contributing to its open-source libraries, ensuring a broad range of features and capabilities. Python is also known for its versatility as it can be used not only for machine learning but also for web development, scientific computing, and data analysis. However, it is important to note that while Python is a popular choice, it is not the only programming language used in machine learning. Other languages, such as R and Java, are also used for machine learning tasks [23].

3.4.2 Libraries

Importing necessary libraries and tools is an essential step when working on any data science project. These libraries provide functionality for common data manipulation, exploration, and deep learning tasks. In this project, we used a number of libraries to preprocess and classify text data as the following:

numpy: NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices. We used this library to work with numerical data in our project.

ktrain: ktrain is a lightweight wrapper for the Keras deep learning library to help simplify the training of neural networks. We used this library to build and train our text classification model.

pandas: Pandas is a library used for data manipulation and analysis. It provides data structures for efficiently storing and manipulating large datasets. We used this library to read in and preprocess our text data.

chardet: Chardet is a Python library used for character encoding detection. We used this library to ensure that our text data is properly encoded before processing.

matplotlib: Matplotlib is a data visualization library used for creating static, animated, and interactive visualizations in Python. We used this library to create visualizations of our data and model performance.

seaborn: Seaborn is a data visualization library based on matplotlib. It provides a high-level interface for creating informative and attractive statistical graphics. We used this library to visualize the distribution of our data.

pickle: Pickle is a Python module used for serializing and de-serializing Python objects. We used this library to save our trained model for future use.

3.4.3 Development environment

3.4.3.1 Google Colab Pro

Since training deep learning models often requires high-performance hardware or can be time-consuming, we selected Google Colab Pro as the platform for our project. Google Colab Pro is a cloud-based service that provides access to a powerful GPU and RAM, which allows us to train our model in a reasonable amount of time.

Specifically, Google Colab Pro provides access to a Tesla K80 GPU, which has 12 GB of GDDR5 VRAM and 4992 CUDA cores. This GPU is suitable for training deep learning models with moderate to high computational requirements.

In addition, **Google Colab Pro** also provides access to TPUs, which are specifically designed for accelerating deep learning computations. TPUs are available for users on a case-by-case basis and require a separate application process.

The TPU offered by Google Colab Pro is the TPU v3-8, which has 8 TPU cores and 64 GB of High Bandwidth Memory (HBM). This TPU is designed for high-throughput deep learning workloads and can accelerate training times by orders of magnitude compared to a traditional GPU.

Overall, the availability of both GPU and TPU on Google Colab Pro made it a suitable platform for our deep learning project, enabling us to train and test our model efficiently and effectively."

Additionally, Google Colab Pro provides a user-friendly interface that allows us to write and execute our code using Jupyter notebooks. This platform also offers other useful features such as version control, collaboration tools, and cloud storage for our data and code.

3.4.3.2 Jupyter notebook

Jupyter notebook is an open-source web application that allows users to create and share documents that contain live code, equations, visualizations, and narrative text. It supports over 40 programming languages, including Python, R, and Julia, making it a versatile tool for data analysis and machine learning [24].

Jupyter notebooks are interactive and allow users to execute code in a step-by-step manner, making it easy to debug and analyze results. They also support the use of Markdown, a markup

language for text formatting, making it easy to create readable and well-structured documents. Jupyter notebooks are widely used in the data science community due to their flexibility, interactivity, and ease of use [25].

3.5 Dataset

In order to train a successful and effective deep learning model, the dataset must be carefully processed, to achieve this, we needed to find a Dataset consisting of samples divided into two classes: "malicious queries" and "non-malicious queries". By using this Dataset, the model can learn to distinguish between the two classes and accurately identify SQL injection attacks.

During our search for a suitable Datasets, we came across the "SQL Injection Dataset" list on the Kaggle platform [26]. We found a list of datasets that were created by “Syed Hussain” that contain Three (03) versions:

- ✓ **SQLi.csv** (723.15 kB) contains **3951** samples with **78%** classified as normal queries and **28%** as malicious queries.
- ✓ **SQLiV2.csv** (3.61 MB) contains **33726** samples with **66%** classified as normal queries and **34%** as malicious queries.
- ✓ **SQLiV3.csv** (2.32 MB) contains **30873** samples with **62%** classified as normal queries and **37%** as malicious queries and **1%** as other.

At first vision we choose the **SQLiV2** Dataset because of the size of samples in it in comparison to others, we trained our model with it in the first time but unfortunately, the results were not satisfied while predicting normal queries. After analyzing the situation, we found that there were no normal queries in the dataset, only free text in place flagged as normal queries.

Because of the big issue in **SQLiV2.csv**, we tried the **SQLiV3.csv** Dataset**.** After reviewing it, we identified certain deficiencies that need to be cleared using some preprocessing steps like:

- ✓ Remove any unnecessary Strings or characters in the SQL statements. We found two commas („) at the end of all the queries, we removed these unnecessary commas using a Register-Expression technique on a python script.
- ✓ Remove not valid empty columns. We need only two valid columns, the **Statement** and the **Label**, in that Dataset we found two empty and not valid columns that were removed using the Office Excel Software.

- ✓ Remove empty and free text Rows. We found many empty and free text rows that were removed using the Office Excel Software.
- ✓ Remove wrong SQL queries. We found not valid SQL statements that were identified and cleared manually.

By applying the above-preprocessed steps, we finally came up with a partitioned version that has **11308 (≈ 50%)** "non-malicious queries" against **11291 (≈ 50%)** "malicious queries".

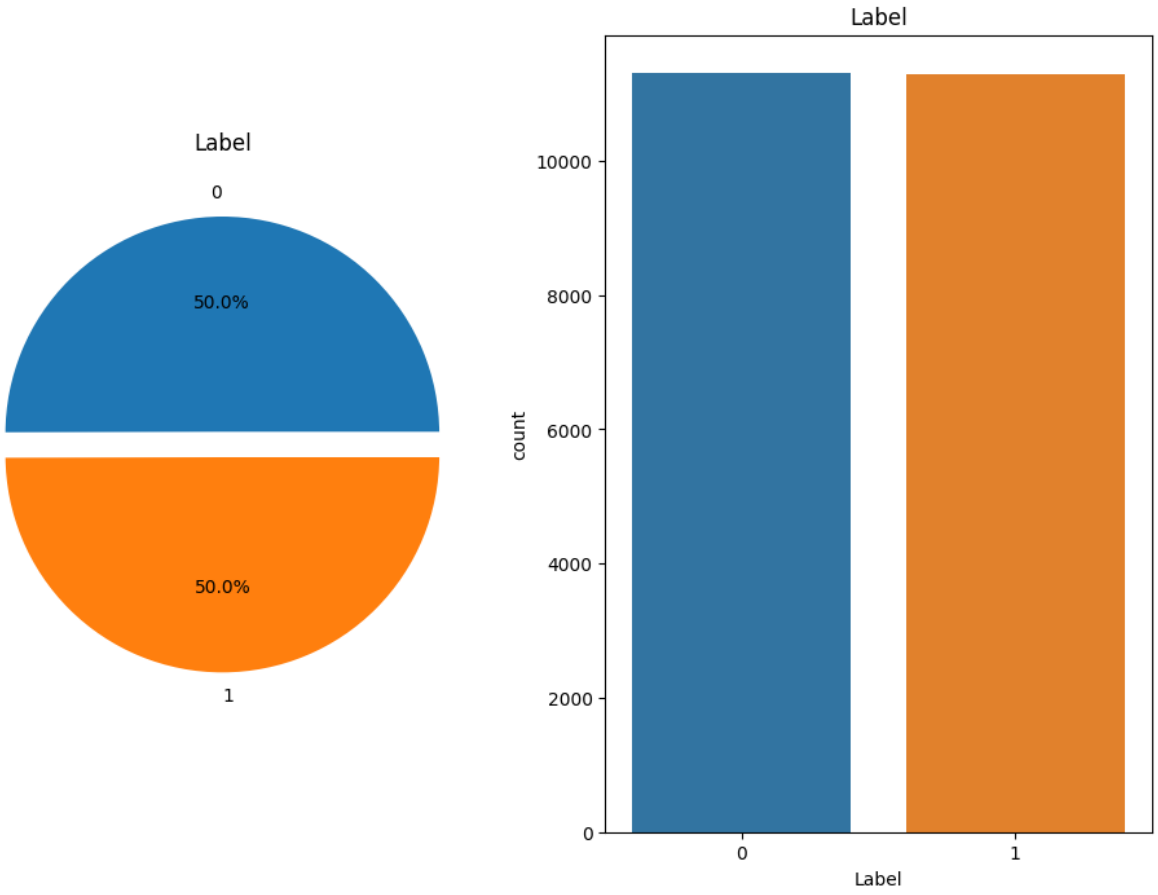

Figure 3.4 Dataset query classes distribution.

After clearing the different deficiencies, we came up with a new valid preprocessed Dataset that has sufficient diversity in both categories and is ready to be used in the next step. We ensured that our deep learning model would only train on preprocessed samples that belong to one of the two classes and could, therefore, learn to differentiate between malicious and non-malicious queries effectively.

3.6 Code and Implementation

The implementation of a deep learning model can be a challenging task, especially when it involves complex architectures and large datasets. In this section, we will present the code and implementation details of our SQL injection detection model based on the BERT architecture.

3.6.1 Split and Preprocess data for the BERT model

```
sentences = df['Sentence'].tolist()

labels = df['Label'].tolist()

(x_train, y_train), (x_test, y_test), preproc =

text.texts_from_array(sentences, labels, preprocess_mode='bert',

maxlen=100, val_pct=0.2, class_names=list(set(labels))

)
```

Figure 3.5 Split data into training and testing sets and preprocess data for BERT model.

In Figure 3.5, the code is splitting the data into training and testing sets and prepares it to be used for training the model. It starts by extracting the "Sentence" and "Label" columns from the pandas dataframe and creating two lists out of them. Next, the "texts_from_array" method from the ktrain library is used to convert the data into arrays that can be used for training a BERT model. The "preprocess_mode" parameter is set to "bert", which means that the data will be preprocessed according to the requirements of the BERT model. "val_pct=0.2" indicates that 20% of the data will be used for validation during training. The "maxlen" parameter is set to 100, which means that the maximum length of a sentence is set to 100 tokens. The "class_names" parameter is set to the unique labels in the training set, which will be used to create a mapping between the label values and their corresponding names. The method returns four variables: "x_train" and "x_test" are the preprocessed arrays of sentences, "y_train" and "y_test" are the label arrays, and "preproc" is a preprocessor object that was used to preprocess the data.

3.6.2 Build the BERT model

```
model = text.text_classifier('bert', (x_train, y_train), preproc=preproc)
```

Figure 3.6 Build BERT model Python code.

In this code line, a BERT model is being built using the text_classifier function from the ktrain library. This function takes the name of the model as the first argument (in this case, "bert"), the training data (x_train, y_train), and the preprocessing object "preproc" as input.

3.6.3 Fine-tuning the BERT model

```
learner = ktrain.get_learner(model=model,
                             train_data=(x_train, y_train),
                             val_data=(x_test, y_test),
                             batch_size=32)
learner.fit_onecycle(lr=2e-5, epochs=4)
```

Figure 3.7 Fine-tuning BERT model Python code.

In the "Train BERT model" section, the model is trained using ktrain's get_learner method and the one-cycle policy, which involves training the model with a learning rate that linearly increases for the first half of the epochs and then linearly decreases for the second half of the epochs.

First, "get_learner" is called, which creates a Learner object for the specified model and training data (x_train, y_train). It also includes validation data (x_test, y_test) to monitor the model's performance during training, and the "batch_size" is set to 32.

The fit_onecycle method is called on the learner object, this method trains the model using the one-cycle policy for a specified “lr” (learning rate) and number of epochs. In this case, the learning rate is set to 2e-5 and the number of epochs is 4.

During the training, the model's loss and accuracy are displayed for each epoch. The goal is to minimize the loss and maximize the accuracy on the validation set to create a well-performing model.

3.6.4 Make predictions with the BERT model

```
predictor = ktrain.get_predictor(learner.model, preproc)

# make predictions

samples = [

"1'; DROP TABLE users;--",

"UPDATE customers SET phone_number = '555-555-5555' WHERE name = 'John Doe'",

"SELECT COUNT(*) FROM users WHERE username = 'admin' OR 1 = 1",

"UPDATE users SET password = 'newpassword' WHERE username = 'admin'"

      ]

prediction = predictor.predict(samples)
```

Figure 3.8 Make predictions with the trained model.

This code block shows how to use the BERT (our trained model) to make predictions on new data. The “get_predictor()” function loads the trained BERT model and pre-processing pipeline, which are necessary to make predictions on new data. The “predict()” function takes a single input and returns the predicted label.

3.7 Choice of hyperparameters

The choice of hyperparameters plays a critical role in the design and performance of our system for detecting SQL injections using deep learning. In this section, we discuss the key hyperparameters we selected and the reasoning behind these choices.

3.7.1 Preprocessing hyperparameters

Max Length: We set the maximum length of input sequences to 100. This value was determined based on the analysis of the dataset, ensuring that most SQL injection statements can be adequately captured within this limit.

Preprocess Mode: We utilized the "bert" preprocess mode, which applies BERT-specific tokenization and formatting to the input text data. This mode is specifically designed for BERT models and helps optimize the preprocessing step, enhancing the model's ability to understand the context and semantics of the text.

3.7.2 Data splitting hyperparameters

Test Size: We partitioned the dataset into training and testing sets using a test size of 0.2 (20%). This split allocates 80% of the data for training and 20% for testing, ensuring a sufficient amount of data for evaluation while preserving a sizable training set.

3.7.3 Model training hyperparameters

Batch Size: We chose a batch size of 32, which determines the number of training samples processed in each iteration. This value strikes a balance between training speed and memory consumption, considering the available computational resources and dataset size.

Learning Rate: We set the learning rate to 2e-5, a value recommended by Google for fine-tuning BERT models. This learning rate choice enables effective convergence during training while minimizing the risk of overshooting the optimal solution.

Number of Epochs: The model was trained for 4 epochs, meaning the entire training dataset was processed four times. This number of epochs allows the model to learn patterns and generalize well to the dataset without over-fitting.

By carefully selecting these Hyperparameters, including the test size, max length, preprocess mode, batch size, learning rate, and epochs, we aimed to optimize the performance of our

system for detecting SQL injections. These choices were based on prior knowledge, best practices and recommendations. The hyperparameters collectively contribute to the effectiveness and accuracy of our model in identifying SQL injection attacks.

3.8 Conclusion

In this chapter, we have described the design and implementation of a system for detecting SQL injections using deep learning. We utilized a dataset of SQL injection attacks and normal queries to train a BERT model for classification. The preprocessing steps involved converting the text data into a format suitable for BERT model input and splitting it into training and testing sets. By sharing the details of our approach, we have provided a comprehensive understanding of how deep learning can be utilized for SQL injection detection.

Chapter 4

Test and Evaluation

4.1 Introduction

The detection of SQL injections using AI techniques, such as machine learning and deep learning algorithms, has gained the interest of many researchers in this field. These techniques have been shown to be effective at identifying SQL injection attacks with high accuracy.

In this chapter, we will discuss the test and evaluation of our model for detecting SQL injection attacks. We will use a variety of metrics, including accuracy, precision, recall, and F1 score. We will also compare the performance of our model to other machine learning algorithms and related works.

4.2 Confusion matrix

The confusion matrix provides a tabular representation of the model's predictions against the actual labels. It allows us to visualize the distribution of true positives, true negatives, false positives, and false negatives, providing valuable insights into the model's performance [27].

	Positive Prediction	**Negative Prediction**
Positive Class	True Positive (TP)	False Negative (FN)
Negative Class	False Positive (FP)	True Negative (TN)

Table 4.1. Confusion Matrix.

TP (True Positives): Correctly predicted positive instances.

TN (True Negatives): Correctly predicted negative instances.

FP (False Positives): Incorrectly predicted positive instances.

FN (False Negatives): Incorrectly predicted negative instances.

4.3 Evaluation metrics for assessing model performance

In the field of machine learning and classification tasks, evaluation metrics play a crucial role in assessing the performance of models. These metrics provide quantitative measures that help us understand the accuracy, effectiveness, and reliability of model predictions. When evaluating the performance of classification models, it is essential to examine the appropriate evaluation metrics that provide insights into their strengths and weaknesses. In this section, we will explore some of the most commonly used evaluation metrics that provide valuable insights into the performance of classification models.

4.3.1 Accuracy

Accuracy is a commonly used evaluation metric that measures the overall correctness of model predictions. It calculates the ratio of correct predictions to the total number of instances. Accuracy provides a general overview of the model's performance across all classes [27].

$$\textbf{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

4.3.2 Precision

Precision focuses on the proportion of correctly identified positive predictions (true positives) out of the total positive predictions made by the model. It helps assess the model's ability to minimize false positives [28].

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

4.3.3 Recall

Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions captured by our model out of the total actual positive instances. It reflects the model's ability to minimize false negatives [28].

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

4.3.4 F1 Score

The F1 score is a combined metric that balances precision and recall. It provides a harmonic mean of these two measures and offers a comprehensive evaluation of the model's performance [29].

$$\textbf{F1 Score} = \frac{2 * Precision * Recall}{Precision + Recall}$$

By analyzing these evaluation metrics, we can gain a deeper understanding of how our classification model performs and identify areas for improvement. These metrics provide valuable insights into the model's strengths and weaknesses, allowing us to make informed decisions to enhance its performance.

4.4 Model performance analysis

In this section, we evaluate the performance of our SQL injection detection model based on the BERT architecture. As already discussed the model was trained on a dataset containing 22,599 samples, consisting of both SQL injections instances and normal SQL queries.

We assess the model's performance using a variety of evaluation metrics, including accuracy, precision, recall, F1 score, and the confusion matrix. These metrics provide valuable insights into the model's ability to correctly classify SQL injection instances and non-malicious queries.

The performance results obtained are as follows:

Accuracy: The model achieved an accuracy of **100%** during training and **99.98%** during validation, indicating its high level of accuracy in classifying the queries.

F1 Score: The F1 score, which considers both precision and recall, also reached an impressive value of **99.98%.** This indicates a balance between correctly identifying normal queries and capturing all actual SQL injection instances.

Precision: The precision of the model, which measures the proportion of true positive predictions out of all positive predictions, was **100%.** This indicates a very low rate of false positives, meaning that the model maintains a high level of confidence in its SQL injection detection predictions.

Recall: The recall of the model, which measures the proportion of true positive predictions out of all actual positive instances, was **99.96%.** This indicates the model's ability to capture nearly all instances of normal queries, resulting in a low rate of false negatives.

The confusion matrix provides a more detailed breakdown of the model's performance:

	Positive Prediction	**Negative Prediction**
Positive Class	2251	1
Negative Class	0	2268

Table 4.2. Confusion Matrix (Classification Results).

The confusion matrix reveals that out of the 2252 positive instances (normal queries), the model correctly identified 2251 instances as positive (true positives) while misclassifying one instance as negative (false negatives). It also correctly classified 2268 out of 2268 negative instances as negative (SQL injections).

These results are consistent with the findings of Srishti Lodha and Atharva Gundawar from the Department of Computer Science and Engineering at Vellore Institute of Technology [30], who made a similar study using the BERT architecture for SQL injection detection. Their research demonstrated comparable performance and highlighted the effectiveness of the BERT model in accurately identifying SQL injection attacks.

Overall, the performance results demonstrate the high accuracy, precision, recall, and F1 score of our SQL injection detection model. These results, in alignment with the work of Srishti Lodha and Atharva Gundawar, underscore the effectiveness of the BERT model for accurately detecting SQL injection attacks while minimizing false positives and false negatives.

4.5 Evaluate the presence of overfitting

It is essential to analyze the loss values of the model during training. By observing the training and validation loss curves, we can gain insights into the model's generalization performance and potential overfitting issues.

The training loss curve represents the loss value computed during the training phase, while the validation loss curve represents the loss value calculated on a separate validation dataset. These curves provide a visual representation of how the model's performance evolves over epochs.

A key indicator of overfitting is when the training loss continues to decrease while the validation loss either stagnates or starts to increase. This suggests that the model is becoming overly specialized to the training data and is failing to generalize well to new, unseen examples.

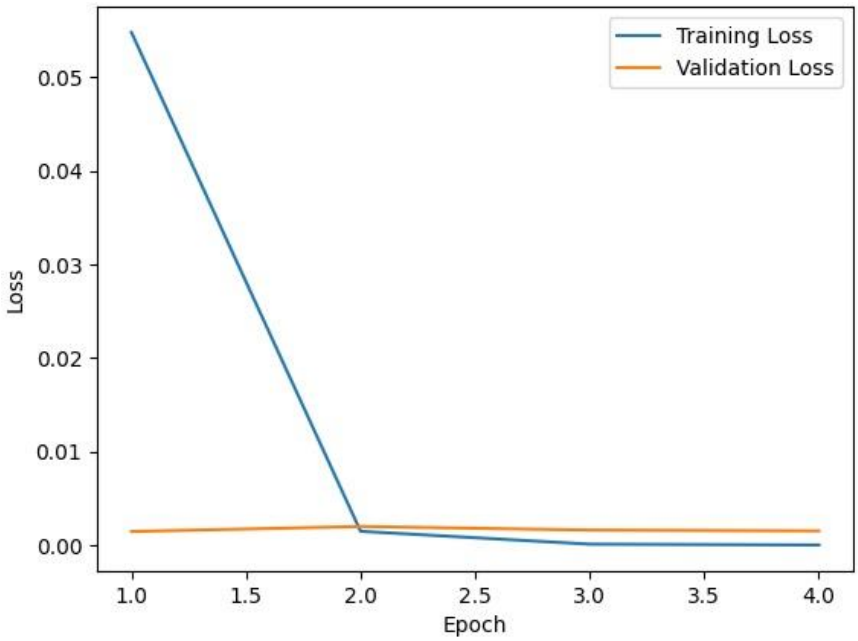

Figure 4.1 Training and validation loss

As shown in Figure 4.1, the training loss consistently decreases over the 4 epochs, indicating that the model is learning from the training data and improving its performance. The validation

loss also shows a decreasing trend, with fluctuations between epochs. However, the general trend is decreasing, which is a positive sign.

The fact that the validation loss generally follows the same decreasing trend as the training loss suggests that the model is not overfitting excessively. It's worth noting that some fluctuations in the validation loss between epochs are normal and can be attributed to random variations in the data or model training process.

4.6 Comparative analysis with other approaches

In this section, we present a comparative analysis of our BERT-based SQL injection detection model with other commonly used approaches available in the literature. While we didn't evaluate the performance of the alternative approaches ourselves, we have collected and compiled information from various sources on their reported performance. By comparing our model with these approaches, we aim to provide insights into the effectiveness of our BERT-based model for SQL injection detection. The comparison is based on commonly used evaluation metrics such as accuracy, precision, recall, and F1 score. The findings of this analysis are summarized in the following table.

Model name	Training Accuracy	Validation Accuracy	Precision	Recall	F1
KNN	100	99.12	98.85	99.52	99.18
SVM	92.78	92.44	90.21	96.36	93.19
BERT (our model)	**100**	**99.98**	**100**	**99.96**	**99.98**

Table 4.3. Comparing the model performances using various metrics (%).

The compared approaches were trained on a dataset of approximately 42,000 data points, while our dataset consisted of 22,599 samples. This disparity arises from the fact that we

obtained the original dataset from the same resource, but we had to perform data cleaning and preprocessing to ensure its quality.

The comparison shows that BERT (Our model) performed the best among the compared models in terms of key evaluation metrics, including accuracy, precision, recall, and F1 score. These results affirm the effectiveness of BERT in detecting SQL injection attacks, highlighting its superior performance in our study.

4.7 Model Performance evaluation on new Data

In this section, we present the results of testing our deep learning model on a new set of data samples. These samples serve as a rigorous assessment of our model's capabilities in handling unseen instances and provide valuable insights into its performance.

	Positive Prediction	**Negative Prediction**
Positive Class	10	0
Negative Class	0	10

Table 4.4. Confusion Matrix (Test Classification Results).

The accuracy of our model on the new data samples stands at a remarkable 100%. This indicates that our model achieved a perfect prediction rate, accurately identifying SQL injections and effectively distinguishing them from normal queries.

Furthermore, the recall score for the positive class (normal queries) is 100%, indicating that our model successfully identified all instances of normal queries present in the new dataset. This showcases its ability to capture the entirety of positive cases and avoid any false negatives.

Similarly, the precision is also 100%. This highlights the model's reliability in identifying SQL injections accurately and minimizing the risk of false positives.

Lastly, the F1 score, which combines precision and recall, also reaches a perfect 100% for the positive class. This score represents the overall balance between accurate detection and

avoiding false positives, emphasizing the model's ability to maintain a high level of performance across both measures.

4.8 Conclusion

In conclusion, our model for detecting SQL injection attacks has shown impressive effectiveness. Through testing and evaluation using various metrics, we have demonstrated its accuracy in identifying SQL injection attacks also the comparison with other models further validates its superior performance.

General Conclusion

This project has made significant contributions to the field of web application security by employing BERT for the detection of SQL injections. The primary contribution of our work lies in demonstrating the effectiveness of BERT in accurately identifying SQL injection attempts in real-time. By exploiting the contextual understanding and semantic representation capabilities of BERT, we have achieved superior performance compared to other machine learning models. Our results provide valuable insights into the application of deep learning techniques for mitigating the major risk of SQL injections on web applications.

Despite the promising results obtained, our work has certain limitations that should be admitted. Firstly, the dataset used for training and evaluation, while comprehensive, may not be large enough to fully exploit the potential of BERT. With larger datasets, BERT's ability to capture complex patterns and nuances could be further enhanced. Secondly, we acknowledge that our experiments were conducted in a controlled environment, and the model was not tested or employed in a real-world scenario.

Based on our analysis of the BERT-based model for SQL injection detection, there are several recommendations and future directions to further enhance its capabilities:

- Explore a larger and more diverse dataset. A larger and more diverse dataset will help the model to learn more about different attack scenarios and improve its performance.
- Investigate the model's performance in real-world scenarios. The model's performance in real-world scenarios should be investigated to understand its limitations. This can be done by deploying the model in a production environment and monitoring its performance.
- Continuously update the model with new attack patterns. The model should be continuously updated with new attack patterns to improve its accuracy in detecting new attacks.
- Expand the scope to detect and classify various types of cyber attacks. The scope of the model can be expanded to detect and classify various types of cyber attacks such as Cross-site scripting (XSS) and Cross-Site Request Forgery (CSRF) using a multi-classification model. This will provide a more comprehensive approach to threat detection and mitigation in cybersecurity.

By addressing these recommendations and future directions, we can advance the field of SQL injection detection and contribute to the development of more effective security solutions.

References

[1] OWASP, "Top 10 Web Application Security Risks," 2021. [Online]. Available: https://owasp.org/www-project-top-ten/. [Accessed 03 June 2023].

[2] OWASP, "SQL Injection," 2023. [Online]. Available: https://owasp.org/www-community/attacks/SQL_Injection. [Accessed 24 Mai 2023].

[3] T. Atefeh , I. Suhaimi and . S. Mohammad, "Web Application Security by SQL Injection DetectionTools," 2012. Available: https://www.researchgate.net/publication/265947554_Web_Application_Security_by_SQL_Injection_DetectionTools

[4] Justin Clarke-Salt, SQL Injection Attacks and Defense, 2nd Edition, Elsevier, Inc, 2012.

[5] OWASP, «Blind SQL Injection,» 2023. . [Online]. Available: https://owasp.org/www-community/attacks/Blind_SQL_Injection. [Accessed 24 Mai 2023]

[6] Dafydd, Stittard Marcos Pinto, The Web Application Hacker's Handbook: Finding and Exploiting Security Flaws, 2nd Edition, Wiley Publishing Inc, 2011.

[7] Positive Technologies, "how-to-prevent-sql-injection-attacks,". [Online]. Available: https://www.ptsecurity.com/ww-en/analytics/knowledge-base/how-to-prevent-sql-injection-attacks. [Accessed 24 Mai 2023]

[8] Mitchell, T. M, Machine Learning. McGraw-Hill, 1997.

[9] Hastie, T., Tibshirani, R., & Friedman, J, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, Springer, 2009

[10] Bishop, C. M, Pattern recognition and machine learning, Springer, 2006.

[11] Sutton, R. S., & Barto, A. G, Reinforcement learning: an introduction, 2nd Edition, MIT Press, 2018.

[12] Goodfellow, I., Bengio, Y., & Courville, A, Deep learning, MIT Press, 2016

[13] Nielsen, M. A, Neural Networks and Deep Learning, Determination Press, 2015.

[14] Graves, A. 2013. Generating sequences with recurrent neural networks. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2013arXiv1308.0850G/abstract, [Accessed 24 Mai 2023]

[15] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y, 2014, Empirical evaluation of gated recurrent neural networks on sequence modeling, [Online]. Available: https://arxiv.org/abs/1412.3555

[16] Sepp Hochreiter, & Jürgen Schmidhuber, Long short-term memory. Neural computation, MIT, 1997, [Online]. Available: https://direct.mit.edu/neco/article-abstract/9/8/1735/6109/Long-Short-Term-Memory

[17] Graves Alex, Supervised sequence labelling with recurrent neural networks, Studies in computational intelligence, Springer 2012 .

[18] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y., Learning phrase representations using RNN encoder-decoder for statistical machine translation, 2014, [Online]. Available: https://arxiv.org/abs/1406.1078

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention is all you need, version 5, 2017. [Online]. Available : https://arxiv.org/abs/1706.03762

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Volume 1, Long and Short Papers , 2019.

[21] Yan Zhang, Huan Zhang, Xuanjing Huang, and Jian Zhao, 2020 ,Benchmarking Neural Network Models for SQL Injection Detection.[Online]. Available : https://www.sciencedirect.com/science/article/abs/pii/S0950705120300332.

[22] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly Media, 2019.

[23] Raschka .S, & Mirjalili. V, Python Machine Learning: Machine Learning and Deep Learning with Python, Packt Publishing, 2019.

[24] Kluyver. T, Ragan-Kelley. B, Pérez. F, Granger. B, Bussonnier, M., Frederic. J, Kluyver, Jupyter Notebooks—a publishing format for reproducible computational workflows, In ELPUB, 2016.

[25] Rule. A, Tabard. A, & Hollan, J. D., Rapid prototyping interactive data visualizations with Jupyter notebooks, IEEE transactions on visualization and computer graphics, 2018.

[26] "SQL injections Dataset on Kaggle," 2023. [Online]. Available: https://www.kaggle.com/datasets/syedsaqlainhussain/sql-injection-dataset. [Accessed 09 Mai 2023]

[27] Sokolova, M., & Lapalme, G., A systematic analysis of performance measures for classification tasks, 2009. [Online]. Available : https://www.sciencedirect.com/science/article/abs/pii/S0306457309000259

[28] D. M. Powers, Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation, Journal of Machine Learning Technologies, 2011.

[29] C. J. Van Rijsbergen, Information Retrieval, 2nd Edition, Butterworths, 1979.

[30] Gundawar, Srishti Lodha and Atharva, SQL Injection and Its Detection Using Machine Learning Algorithms, 2023. [Online]. Available : https://link.springer.com/chapter/10.1007/978-3-031-28975-0_1