Universiti Teknologi MARA

CROSS-SITE DETECTION SYSTEM USING SUPPORT VECTOR MACHINE

AHMAD A'LIMMUDDIN BIN BAHRIM

Thesis submitted in fulfilment of the requirements
for Bachelor of Computer Science (Hons.)
Faculty of Computer and Mathematical Science

July 2023

ACKNOWLEDGMENT

Alhamdulillah, praise and thanks to Allah because of His Almighty and His greatest blessings, I was able to finish this research by the deadline. My supervisor, Sir Muhammad Atif Ramlan, has been a great help to me during this research, and I would want to express my profound gratitude for that. He was essential to my ability to do this assignment successfully; without him, I could not have.

I would like to express my sincere gratitude to my professor, Madam Ummu Fatihah Mohd Bahrin, for her help and expertise in the area of CSP650, both of which have been extremely helpful to me in my research. A special mention should also be made to my adored parents for their constant support and inspiration during this effort. I will always be appreciative of their support and kindness.

Last but not least, I would want to thank all of my beloved friends and those who have participated in this adventure and have given me their unflagging support and encouragement. This project would not have been possible without them.

ABSTRACT

A form of security issue called cross-site scripting (XSS) enables attackers to insert malicious code into a website. When a user accesses the website, the malicious code may steal personal data or carry out other undesirable actions. XSS attacks can be classified as stored, reflected, or DOM-based. With the help of machine learning techniques like Support Vector Machines (SVM), these attacks, which are frequent, can be stopped. A cross-site detection system for XSS scripting was created in this work utilising the Support Vector Machine (SVM) technique. Support Vector Machine (SVM) is a technique used to determine whether XSS scripts have been implanted in a website or not. Six different research approaches, including a preliminary study, requirement analysis, data gathering, design, implementation, evaluation, and documentation, were used to construct this system efficiently. The system's stated objectives could be successfully attained at the end of the study thanks to the tight alignment of these approaches with those goals. Next, the dataset used for this study is dataset named "Cross site scripting XSS dataset for Deep learning" can be download from website online which is Kaggle contributed by Syed Saqlain Hussain Shah. The dataset contains Cross site scripting attack (XSS) data along with benign data. The research is significant in addressing the serious threats posed by cross-site assaults to the security and integrity of web systems, and in contributing to the development of effective detection and mitigation strategies.

Table of Contents

ACKNOWLEDGMENT 5

ABSTRACT 6

CHAPTER ONE 11

INTRODUCTION 11

1.1 Background Study 11

1.2 Problem Statement 12

1.3 Question 13

1.4 Objectives 13

1.5 Project Scopes 14

1.6 Project Significance 15

1.7 Overview of Research Framework 16

1.8 Conclusion 16

CHAPTER TWO 17

LITERITURE REVIEW 17

2.1 Introduction 17

2.2 Overview of Cross-site Detection 18

2.2.1 Cross-site Scripting (XSS) 18

2.2.2 Cross-site Scripting Affect 19

2.2.4 Benefits of Cross-site Detection 20

2.3 Machine Learning 21

2.3.1 SVM Algorithm and How Does It Works 22

2.4 Implementation SVM Algorithm in Various Problem 23

2.5 Similar works 29

2.6 The Implication of Literature Review 34

2.7 Conclusion 35

CHAPTER THREE 36

RESEARCH METHODOLOGY 36

3.1 Overview of Research Methodology 36

3.1.1 Detailed of Research Framework 36

3.2 Preliminary Phase 39

3.2.1 Literature Study 40

3.3 Requirement Analysis 40

3.3.1 Knowledge Acquisition 40
3.4 Data Preprossesing 41
3.4.1 Data Collection 41
3.4.3 Data Cleaning 42
3.5 Design Phase 43
3.5.1 System Architecture 44
3.5.3 User Interface 47
3.5.4 Pseudocode of Selected Algorithm 47
3.5.5 Prototype Implementation 49
3.6 Evaluation Phase 50
3.7 Documentation 51
3.8 Gantt Chart 52
3.9 Conclusion 52
CHAPTER FOUR 53
RESULTS AND FINDINGS 53
4.1 Conceptual Framework 53
4.2 Result for Objective 1 54
4.2.1 Text Preprocess 54
4.2.2 Data Splitting 55
4.3 Result for Objective 2 57
4.3.1 SVM Implementation 57
4.3.2 Prototype Interface 60
4.4 Result for Objective 3 62
4.4.1 SVM Model Evaluation 62
4.4.2 Discussion ..... Error! Bookmark not defined.
4.5 Conclusion 68
CHAPTER FIVE 69
CONCLUSION 69
5.1 Project Summary 69
5.2 Project Limitations 70
5.3 Future Recommendations 70
5.4 Project Contributions 71
5.5 Conclusion 72
References 74