A hybrid approach on phishing URL Detection using Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU)

B.A.S. Dilhara
MS20907402
M.Sc. in IT
Specialized in Cyber Security

Supervisor: Dr Dharshana Kasthurirathna

December 2021

Faculty of Graduate Studies & Research
Sri Lanka Institute of Information Technology

Declaration

This research project is all my work and it has not been copied in part or whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, and the internet, etc.) has been acknowledged within the main report to an item in the reference. This research project was submitted in partial fulfillment of the requirement of the award of **Master of Information Technology (Specialized) in Cyber Security**.

Copyright Acknowledgement:

I acknowledged that the joint copyright of this project report belongs to the Sri Lanka Institute of Information Technology and myself.

Certified by:

Signature: ………………………………

Date: ………………………………

Name of Supervisor: Dr. Dharshana Kasthurirathna

Signature: …………………………………….

Date: …………………………………….

Acknowledgement

I would like to convey my heartfelt gratitude to my research supervisor, Dr Dharshana Kasthurirathna, Assistant Professor at Sri Lankan Institute of Information Technology for providing me continuous guidance and supervision throughout the research.

In addition, I would also like to convey my heartfelt gratitude to all of my office colleagues. It is their kind help and support that have made my study and my work very exciting.

Most importantly I would like to convey my gratefulness to my mother and my sister for their love, understanding and continuous love and support. They have been always been my pillars of success by being there for me whenever I needed it most. Additionally, I also would like to thank everyone who gave me helping hand in many ways to complete my master's degree.

Finally, this thesis is dedicated to my beloved late father who guided me in the correct path to become the person who I am today.

Table of Contents

Declaration 2

Acknowledgement 3

Table of Contents 4

List of Figures 7

List of Tables 9

Abstract 10

Chapter 1 Introduction 11

1.1 Background 11

1.2 Significance of the Study 13

1.3 Problem Statement (Definition) 13

1.4 Research Objectives and Research Questions 14

1.4.1 Research Objectives 14

1.4.2 Research Questions 14

Chapter 2 Literature Review 15

2.1 Traditional Methods 15

2.2 Machine Learning Approach 16

2.3 Deep Learning Approach 19

2.4 Machine Learning Approach using Hybrid Models 22

2.5 Deep Learning Approach using Hybrid Models 24

Chapter 3 Methodology 29

3.1 Dataset Selection 29

3.2 Deep Learning Models Used 31

3.2.1 Recurrent Neural Network (RNN) 32

3.2.2 Long Short-Term Memory (LSTM) 33

3.2.3 Gated Recurrent Units (GRU) 36

3.2.4 1D Convolutional Neural Network (CNN 1D) 37

3.3 Ensemble Approach 38

3.4 Sequential Model and Functional API models 39

3.5 Software, tools, languages and libraries used 39

3.5.1 Python Language 39
3.5.2 Frameworks and libraries 40
3.5.3 Jupyter Notebook 41
3.6 Proposed Approach 41
3.7 Design and Implementation 42
3.7.1 Model overview of Convolutional 1D model 43
3.7.2 Model overview of Long Short Term memory model 44
3.7.3 Model overview of Gated Recurrent Unit 45
3.7.4 Model Overview of LSTM – LSTM model 46
3.7.5 Model Overview of GRU – LSTM model 48
3.7.6 Model Overview of Bi (LSTM) – LSTM model 50
3.7.7 Model Overview of Bi (GRU) – LSTM model 52
3.7.8 Dataset Preprocessing 54
3.8 Training of the proposed models 55
3.9 Testing of the proposed models 58
3.10 Evaluation of the proposed deep learning models 59
3.10.1 Confusion Matrix 60
3.10.2 Accuracy 61
3.10.3 Precision 61
3.10.4 Recall 62
3.10.5 F1 – Score 62
Chapter 4 Results and discussion 63
4.1 Model Predictions 63
4.1.1 CNN (1D) 63
4.1.2 LSTM 65
4.1.3 GRU 67
4.1.4 LSTM -LSTM 69
4.1.1 GRU -LSTM 71
4.1.2 Bi (LSTM) - LSTM 73
4.1.3 Bi (GRU) – LSTM 75
4.2 Results Comparison 78
Chapter 5 Conclusion 85

5.1 Limitations and drawbacks 85
5.2 Future Work 86
5.3 Conclusion about the research problem 86
Reference 88
APPENDIX 93
Appendix 1: Data Preprocessing Code 93
Appendix 2: CNN (1D) Model 94
Appendix 3: LSTM Model 97
Appendix 4: GRU Model 99
Appendix 5: LSTM – LSTM Model 101
Appendix 6: GRU – LSTM Model 104
Appendix 7: Bi (LSTM) – LSTM Model 107
Appendix 8: Bi (GRU) – LSTM Model 110

List of Figures

Figure 3.1 Distribution of the Dataset 30
Figure 3.2 Algorithm for feature extraction process 31
Figure 3.3 URL Structure Breakdown 31
Figure 3.4 RNN Cell Hidden State Calculation [41] 32
Figure 3.5 Vector Transformations with Tanh 33
Figure 3.6 LSTM and its Cell Operation 34
Figure 3.7: Forget gate, input gate, output gate and cell state [41]. 35
Figure 3.8 GRU Cell 36
Figure 3.9 A Sample 1D – CNN Architecture 37
Figure 3.10 Structure of the model proposed for non-hybrid models 42
Figure 3.11 Structure of the model proposed of the hybrid models 42
Figure 3.12 CNN (1D) proposed Architecture 43
Figure 3.13 LSTM model flow 44
Figure 3.14 LSTM Model Architecture 44
Figure 3.15 GRU Model Architecture 45
Figure 3.16 GRU Model Flow 45
Figure 3.17 LSTM - LSTM Model Architecture 46
Figure 3.18 LSTM - LSTM Model Flow 47
Figure 3.19 GRU - LSTM Model Architecture 48
Figure 3.20 GRU - LSTM Model Flow 49
Figure 3.21 Bi (LSTM) -LSTM Model Architecture 50
Figure 3.22 Bi (LSTM) - LSTM Model Flow 51
Figure 3.23 Bi (GRU) -LSTM Model Architecture 52
Figure 3.24 Bi (GRU) -LSTM model flow 53
Figure 3.25 Dataset pre-processing and saving to a physical file as NumPy arrays 55
Figure 3.26 A visualization of the splits and training flow 57
Figure 3.27 Iterative process of evaluating the deep learning model 59
Figure 4.1 Variation of Accuracy and Validation Accuracy over the number of epochs in CNN (1D) model 63
Figure 4.2 Variation of Loss and Validation Loss over the number of epochs in CNN (1D) model 64
Figure 4.3Confusion Matrix of proposed CNN (1D) model 64
Figure 4.4 Variation of Accuracy and Validation Accuracy over the number of epochs in LSTM model ... 65
Figure 4.5 Variation of Loss and Validation Loss over the number of epochs in LSTM model 66
Figure 4.6 Confusion Matrix of proposed LSTM model 66
Figure 4.7 Variation of Accuracy and Validation Accuracy over the number of epochs in GRU model 67
Figure 4.8 Variation of Loss and Validation Loss over the number of epochs in GRU model 68
Figure 4.9 Confusion Matrix of proposed GRU model 69
Figure 4.10 Variation of Accuracy and Validation Accuracy over the number of epochs in LSTM – LSTM model 70

Figure 4.11 Variation of Loss and Validation Loss over the number of epochs in LSTM - LSTM model 70
Figure 4.12 Confusion Matrix of proposed LSTM - LSTM model 71
Figure 4.13 Variation of Accuracy and Validation Accuracy over the number of epochs in GRU – LSTM model 72
Figure 4.14 Variation of Loss and Validation Loss over the number of epochs in GRU - LSTM model 72
Figure 4.15 Confusion Matrix of proposed GRU - LSTM model 73
Figure 4.16 Variation of Accuracy and Validation Accuracy over the number of epochs in bi (LSTM) – LSTM model 74
Figure 4.17 Variation of Loss and Validation Loss over the number of epochs in bi (LSTM) – LSTM model 74
Figure 4.18 Confusion Matrix of proposed Bi (LSTM) - LSTM model 75
Figure 4.19 Variation of Accuracy and Validation Accuracy over the number of epochs in Bi (GRU) – LSTM model 76
Figure 4.20 Variation of Loss and Validation Loss over the number of epochs in bi (GRU) – LSTM model 76
Figure 4.21 Confusion Matrix of proposed bi (GRU) - LSTM model. 77
Figure 4.22 : Graphical Representation of TNR, TPR, FPR, and FNR values 80
Figure 4.23 Figure 48: A Bar Graph showing the values of Accuracy, Validation Accuracy, Precision, Recall, and 83

List of Tables

Table 3.1: Confusion Matrix 60
Table 4.1: Results obtained from the evaluation of the proposed Deep Learning models 78
Table 4.2 Ranking of proposed models based on the results of FPR, FNR, TPR, and TNR 79
Table 4.3:Ranking of proposed models based on the results of Accuracy, Validation Accuracy, Precision, Recall and F1 Score 82

Abstract

Phishing is one of the oldest types of cyber-attack which mostly comes in the form of camouflaged URLs to delude the users in order to get their personal information for malevolent purposes of the attacker. In addition, it is one of the easiest ways of inducing people into disclosing their personal credentials including credit card details. Since people use web applications on a daily basis, most phishing attacks comes up as fake websites pretending to mimic a trustworthy website. Moreover, emails are being used by the attackers to send the phishing website URL (Uniform Resource Locator) to the victim. Such type of URLs is termed as malicious URLs and most phishing attackers use them for successful data breaches. Therefore, it is a necessity to filter up, which URLs are benign, and which are malicious. In order to determine these factors, the concepts including traditional mechanisms used for URL detection, the drawbacks that those mechanisms had, and machine learning approaches used by different authors and their novelty approaches for effective detection are reviewed through this paper. Moreover, this will be focusing on cumulative deep learning approaches to build up hybrid deep learning models. Furthermore, this study proposes 4 hybrid deep learning models namely GRU-LSTM, LSTM-LSTM, bidirectional (GRU)-LSTM, and bidirectional (LSTM)-LSTM. In addition, the study also proposes 3 non hybrid deep learning models namely CNN(1D), LSTM and GRU. Hence, the main objective of this research is to provide a new insight to the hybrid deep learning approaches in URL detection by evaluating their accuracy, precision, recall and f1 score. In conclusion, this research recognizes Bi (GRU) – LSTM as the best mechanism to join hybrid models to detect phishing URLs and classify them as malicious or benign.