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

List of Figures

List of Tables

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.