การป้องกันการตกเป็นเหยื่อจากการฟิชชิงผ่านทางที่อยู่เว็บไซต์โดยวิธีการเรียนรู้ของเครื่อง

ณิชารีย์ อิ๋วสกุล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
บัณฑิตศึกษา สาขาเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีไทย-ญี่ปุ่น
ปีการศึกษา 2566

MACHINE-LEARNED DEFENSE MECHANISM AGAINST PHISHING URL EXPLOITATION FOR USER PROTECTION

Nicharee Ewasakul

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Information Technology
Graduate Studies
Thai-Nichi Institute of Technology
Academic Year 2023

หัวข้อวิทยานิพนธ์	การป้องกันการตกเป็นเหยื่อจากการฟิชชิงผ่านที่อยู่เว็บไซต์โดยวิธีการเรียนรู้ของเครื่อง
โดย	ณิชารีย์ อิ๋วสกุล
สาขาวิชา	เทคโนโลยีสารสนเทศ
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ดร.ทองทัต โอฬารศิริกุล

บัณฑิตศึกษา สถาบันเทคโนโลยีไทย-ญี่ปุ่น อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... รองอธิการบดีฝ่ายวิชาการ
(รองศาสตราจารย์ ดร.วรากร ศรีเชวงทรัพย์)
วันที่.....เดือน.....พ.ศ.....

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ณัฐสุดา เกาทัณฑ์ทอง)

..... กรรมการ
(ดร.สะพรั่งสิทธิ์ มฤทุสาธร)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ฐิติพร เลิศรัตน์เดชากุล)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์
(ดร.ทองทัต โอฬารศิริกุล)

ณิชารีย์ อิ๋วสกุล : การป้องกันการตกเป็นเหยื่อจากการฟิชชิงผ่านทางที่อยู่เว็บไซต์ โดยวิธีการเรียนรู้ของเครื่อง. อาจารย์ที่ปรึกษา : ดร.ทองทัต โอฬารศิริกุล, 53 หน้า.

ฟิชชิงเป็นการโจมตีทางไซเบอร์ที่ผู้โจมตีทำการโจมตีโดยการหลอกลวงบุคคล หรือองค์กรให้เปิดเผยข้อมูลสำคัญ เช่น รหัสผ่าน หมายเลขบัตรเครดิต หรือข้อมูลส่วนตัว โดยการโจมตีแบบฟิชชิงผ่านทางที่อยู่เว็บไซต์ เป็นวิธีการฟิชชิงที่สามารถพบได้บ่อย โดยผู้โจมตีทำการสร้างที่อยู่เว็บไซต์ปลอมที่คล้ายกับ ที่อยู่เว็บไซต์จริงขององค์กรต่าง ๆ ที่ผู้โจมตีใช้อ้างถึง เพื่อหลอกให้ผู้ใช้คลิกและทำการส่งข้อมูลสำคัญไปยังฝั่งของผู้โจมตี โดยเทคนิคที่ใช้ในการป้องกันการโจมตีแบบฟิชชิงผ่านทางที่อยู่เว็บไซต์แบบดั้งเดิม อย่างเช่น การสร้างฐานข้อมูลที่รวบรวมรายชื่อเว็บไซต์ที่ถูกบล็อกและรายชื่อเว็บไซต์ที่ได้รับอนุญาตให้เข้าถึงได้นั้น มีประสิทธิภาพในการตรวจจับ ได้เพียงที่อยู่เว็บไซต์อันตรายที่รู้จักแล้วเท่านั้น จึงทำให้เกิดข้อจำกัดในการตรวจสอบที่อยู่เว็บไซต์อันตราย (Malicious URL) ใหม่ ๆ ด้วยเหตุนี้ เทคนิคการเรียนรู้ของเครื่องจึงถูกนำมาใช้ในงานตรวจสอบที่อยู่เว็บไซต์อันตรายอย่างแพร่หลายในปัจจุบัน โดยการเลือกใช้คุณลักษณะที่เหมาะสม คือ หนึ่งในปัจจัยที่ช่วยส่งเสริมประสิทธิภาพของแบบจำลองการเรียนรู้ในการจำแนกที่อยู่เว็บไซต์อันตราย (Malicious URL) ซึ่งคุณลักษณะที่ถูกสกัดจากที่อยู่เว็บไซต์นั้นสามารถแบ่งออกได้เป็นสามกลุ่มหลัก ๆ คือ กลุ่มคุณลักษณะทางภาษา (Lexical feature) กลุ่มคุณลักษณะทางเนื้อหา (Content-based feature) และ กลุ่มคุณลักษณะทางเครือข่าย (Network-based feature) โดยงานวิจัยฉบับนี้เป็นการศึกษาเกี่ยวกับลักษณะของกลุ่มคุณลักษณะทางภาษา และ กลุ่มคุณลักษณะทางเนื้อหา ซึ่งเป็นกลุ่มคุณลักษณะที่มีความเกี่ยวเนื่องกับการวิเคราะห์ถึงลักษณะ โครงสร้างทางภาษา และ ทำการทดลองโดยการใช้เทคนิค Recursive feature elimination (RFE) ในการคัดเลือกคุณลักษณะสำคัญของแต่ละกลุ่ม และทำการสร้างโมเดล 2 โมเดล โดยใช้อัลกอริทึมป่าไม้แบบสุ่ม (Random forest) เพื่อเปรียบเทียบประสิทธิภาพในการส่งเสริมการเรียนรู้ในการทำนายที่อยู่เว็บไซต์อันตราย (Malicious URL) ให้แก่แบบจำลองการเรียนรู้ของเครื่อง โดยผลการทดลองแสดงให้เห็นว่า โมเดลที่ใช้คุณลักษณะสำคัญจากกลุ่มคุณลักษณะทางเนื้อหา (Content-based feature) นั้นมีความแม่นยำอยู่ที่ 98% ส่วนโมเดลที่ใช้คุณลักษณะจากกลุ่มคุณลักษณะทางภาษา (Lexical feature) นั้น มีความแม่นยำเพียง 67% เท่านั้น

บัณฑิตศึกษา ลายมือชื่อนักศึกษา .....

สาขาวิชา เทคโนโลยีสารสนเทศ ลายมือชื่ออาจารย์ที่ปรึกษา .....

ปีการศึกษา 2566

NICHAREE EWASAKUL : MACHINE-LEARNED DEFENSE MECHANISM AGAINST PHISHING URL EXPLOITATION FOR USER PROTECTION. ADVISOR : DR. THONGTAT ORANSIRIKUL, 53 PP.

Phishing is a cyberattack where attackers deceive individuals or organizations into providing sensitive information such as passwords, credit card numbers, or personal data. URL phishing is a prevalent method where attackers create fake URLs that closely resemble legitimate ones, tricking users into clicking and potentially compromising their sensitive information. Traditional defense mechanisms like blacklists and whitelists are effective at detecting known malicious URLs but struggle with identifying new ones. Consequently, machine learning has emerged as a robust approach for phishing detection. The effectiveness of machine learning models depends on the features used for training. Features extracted from URLs can be categorized into three main types: lexical-based, content-based, and network-based. Our research aims to determine which feature types enhance the accuracy and effectiveness of machine learning in detecting malicious URLs.

In this study, we focus on comparing the effectiveness of two feature groups: lexical-based and content-based features, both of which involve analyzing the textual properties of URLs. We employ Recursive Feature Elimination (RFE) to select the top 10 features from each group, ensuring that only the most relevant features are utilized. Subsequently, we apply a random forest model to classify the URLs based on these selected features. Our results demonstrate that the 10 features chosen from the content-based group achieve the highest accuracy, reaching up to 98%, while lexical-based features achieve only 67%. This significant finding highlights the potential of content-based features in enhancing the performance of machine learning models for phishing detection.

Graduate Studies Student's Signature.....

Field of Study Information Technology Advisor's Signature.....

Academic Year 2023

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลือจากอาจารย์ที่ปรึกษา ดร.ทองทัต โอฬารศิริกุล ผู้ให้ความรู้ คำชี้แนะและให้ความช่วยเหลือทุกอย่างตลอดการทำวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอกราบขอบพระคุณไว้ ณ โอกาสนี้

ขอกราบขอบพระคุณประธาน และคณะกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.ณัฐสุดา เกาทัณฑ์ทอง ดร. สะพรั่งสิทธิ์ มฤทุสาธร และผู้ช่วยศาสตราจารย์ ดร. ฐิติพร เลิศรัตน์เดชากุล ผู้ให้คำแนะนำ และข้อแก้ไขที่ควรปรับปรุงจนวิทยานิพนธ์ฉบับนี้สมบูรณ์

ขอกราบขอบพระคุณคณาจารย์ สถาบันเทคโนโลยีไทย-ญี่ปุ่น ผู้ประสิทธิ์ประสาทวิชาความรู้ตลอดจนเจ้าหน้าที่ผู้อำนวยความสะดวก และช่วยเหลือทุกอย่าง และขอขอบพระคุณครอบครัว และเพื่อน ๆ ที่ให้คำแนะนำ และกำลังใจมาโดยตลอด

สุดท้ายนี้ผู้วิจัยหวังเป็นอย่างยิ่งว่า วิทยานิพนธ์ฉบับนี้จะเป็นประโยชน์แก่ผู้อื่นต่อไป

ณิชารีย์ อิ่วสกุล

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง
บทคัดย่อภาษาอังกฤษ ..... จ
กิตติกรรมประกาศ ..... ฉ
สารบัญ ..... ช
สารบัญตาราง ..... ฌ
สารบัญรูป ..... ญ

บทที่

1 บทนำ 1
1.1 ความเป็นมาและความสำคัญของปัญหา 1
1.2 วัตถุประสงค์ของการศึกษา 3
1.3 ขอบเขตการวิจัย 3

2 เอกสารและงานวิจัยที่เกี่ยวข้อง 4
2.1 ฟิชชิง (Phishing) 4
2.2 องค์ประกอบของที่อยู่เว็บไซต์ 5
2.3 การเรียนรู้ของเครื่อง (Machine Learning) 6
2.4 คุณลักษณะ (Features) 7
2.5 งานวิจัยที่เกี่ยวข้อง 9

3 วิธีการดำเนินงาน 13
3.1 การจัดเก็บข้อมูล (Data Collection) 14
3.2 การเตรียมข้อมูล (Data Preprocessing) 15
3.3 การสร้างโมเดล (Modeling) 21
3.4 การประเมินผล (Evaluation) 21

สารบัญ (ต่อ)

บทที่		หน้า
4	ผลการวิจัย	24
5	ข้อสรุปและเสนอแนะ	27
	5.1 การสรุปผลและอภิปรายผลการวิจัย	28
	5.2 ข้อเสนอแนะ	29
บรรณานุกรม		30
ภาคผนวก		34
	ภาคผนวก ก. เอกสารประกอบการนำเสนอการสอบป้องกันวิทยานิพนธ์	35
ประวัติย่อผู้วิจัย		53

สารบัญตาราง

ตาราง หน้า

2.1 ตัวอย่างคุณลักษณะทางภาษา 7
2.2 ตัวอย่างคุณลักษณะทางเนื้อหา 8
2.3 ตารางสรุปงานวิจัยที่เกี่ยวข้อง 11
3.1 คุณลักษณะที่ปรากฏในชุดข้อมูล 14
3.2 คุณลักษณะทางภาษาที่ถูกสกัดจากชุดข้อมูล 16
3.3 คุณลักษณะทางเนื้อหาที่ถูกสกัดจากชุดข้อมูล 17
3.4 คุณลักษณะที่สำคัญ 10 อันดับแรก ของกลุ่มคุณลักษณะทางภาษา 19
3.5 คุณลักษณะที่สำคัญ 10 อันดับแรก ของกลุ่มคุณลักษณะทางเนื้อหา 20
3.6 ตาราง Confusion Matrix 22
4.1 ผลการทำนายที่อยู่เว็บไซต์ของแต่ละกลุ่มคุณลักษณะ 25
4.2 ตารางสรุปผลการทดลอง 26

สารบัญรูป

รูป		หน้า
2.1	องค์ประกอบที่อยู่เว็บไซต์	5
3.1	ขั้นตอนการดำเนินงานวิจัย	13

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การพัฒนาของเทคโนโลยีและการสื่อสารที่เติบโตขึ้นอย่างรวดเร็วจากในอดีตจนถึงปัจจุบันนั้น นอกจากจะช่วยให้ผู้คนทั่วโลกสามารถติดต่อสื่อสารถึงกันได้อย่างสะดวก รวดเร็ว และมีประสิทธิภาพมากยิ่งขึ้นแล้ว ความก้าวหน้าของเทคโนโลยีก็ยังเป็นช่องทางสำหรับผู้ที่ไม่ประสงค์ดีใช้สำหรับการคุกคาม และโจมตีบุคคล หรือ องค์กร ได้อีกด้วย ซึ่งในปัจจุบันภัยคุกคามทางไซเบอร์นั้นกลายเป็นปัญหาสังคมอย่างนึงที่ต้องได้รับการจัดการแก้ไขปัญหาอย่างเร่งด่วน เนื่องจากส่งผลกระทบอย่างรุนแรงต่อความปลอดภัยของข้อมูลทั้งในระดับบุคคล และองค์กร

โดยรูปแบบการโจมตีทางไซเบอร์ที่สามารถพบได้บ่อยที่สุด และเป็นที่รู้จักอย่างกว้างขวางในปัจจุบัน คือ รูปแบบการโจมตีแบบฟิชชิง (Phishing) ฟิชชิง คือ รูปแบบการโจมตีที่ใช้วิธีการหลอกล่อผู้ใช้งาน ให้กระทำกิจกรรมที่เสี่ยงต่อความปลอดภัยของข้อมูลส่วนบุคคล ซึ่งมีเป้าหมายเป็นข้อมูลส่วนบุคคล เช่น ชื่อผู้ใช้งาน รหัสผ่าน หรือ หมายเลขบัตรเครดิต เพื่อนำไปสู่การขโมยทรัพย์สิน หรือสร้างความเสียหายแก่ผู้ใช้งานที่ถูกโจมตี โดยผู้โจมตีจะใช้วิธีการแอบอ้างเป็นบุคคล หรือ องค์กรที่น่าเชื่อถือ ทำการส่งข้อความ โดยแนบที่อยู่เว็บไซต์ หรือ ไฟล์ที่มีมัลแวร์ (Malware) ซ่อนอยู่ ไปทางเบอร์โทรศัพท์ หรืออีเมล์ของผู้ใช้งานที่เป็นเป้าหมายในการโจมตี หรือจากการค้นหาที่อยู่เว็บไซต์ ซึ่งหากผู้ใช้งานทำการกดไปยังที่อยู่เว็บไซต์ต้องสงสัยที่แนบมานั้นและทำการเข้าไปกรอกข้อมูลต่าง ๆ ก็จะทำให้ข้อมูลต่าง ๆ เหล่านั้น ถูกขโมยไปนอกจากนี้อาจก่อให้เกิดความเสียหายแก่คอมพิวเตอร์และอุปกรณ์ ของผู้ใช้งานได้อีกด้วย

ในปี 2565 ที่ผ่านมา แคสเปอร์สกี้ รายงานว่าสามารถตรวจพบความพยายามในการโจมตีด้วยรูปแบบฟิชชิง จำนวน 6,283,745 ครั้งในประเทศไทย ซึ่งติดอับดับ 3 ในอาเซียน [1] และ วันที่ 16 มิถุนายน 2566 มีการรายงานจากทางสำนักงานตำรวจแห่งชาติ แผนกกองบัญชาการตำรวจสืบสวนอาชญากรรมทางเทคโนโลยี ถึงยอดสถิติจำนวนการแจ้งความออนไลน์เกี่ยวกับคดีอาชญากรรมทางออนไลน์ ผ่านระบบการแจ้งความของสำนักงานตำรวจแห่งชาติ เมื่อวันที่ 1 มีนาคม 2566 ถึง 31 พฤษภาคม 2566 โดยมีจำนวนคดีทั้งหมด 296,243 คดี คิดเป็นมูลค่าความเสียหายประมาน 40,000 ล้านบาท และได้มีการเปิดเผยต่ออีกว่า เหยื่อในปัจจุบันนั้น ไม่ใช่กลุ่มผู้สูงอายุ แต่มักเป็นกลุ่มคนในช่วงวัยทำงาน ที่ใช้เทคโนโลยีอยู่เป็นประจำ[2] ซึ่งถึงแม้จะมีการแจ้งเตือนต่าง ๆ ทั้งจากภาครัฐและเอกชน ตลอดจนการให้ความรู้และวิธีการป้องกันภัยจากทางไซเบอร์แล้วก็ตาม ก็ยังคงมีผู้เสียหาย

อีกเป็นจำนวนมาก ซึ่งสะท้อนให้เห็นว่าภัยคุกคามทางไซเบอร์ยังคงเป็นหนึ่งในปัญหาสำคัญของสังคมที่จะต้องเร่งศึกษาเพื่อหาแนวทางในการป้องกันให้กับผู้ใช้งานอินเทอร์เน็ต

โดยการตรวจสอบที่อยู่เว็บไซต์ ถือเป็นอีกหนึ่งวิธีที่สามารถช่วยป้องกันและลดความเสี่ยงในการถูกคุกคามด้วยวิธีฟิชชิงผ่านทางที่อยู่เว็บไซต์ได้ โดยจากการศึกษาพบว่าการตรวจสอบที่อยู่เว็บไซต์นั้น ด้วยวิธีการแบบดั้งเดิม สามารถทำได้ด้วยวิธีการจัดทำฐานข้อมูลแบล็คลิสท์ที่อยู่เว็บไซต์ (Blacklist-Based Detection) และ วิธีการฮิวริสติก (Heuristic method) ซึ่ง 2 วิธีการดังกล่าว มีประสิทธิภาพในการตรวจสอบที่อยู่เว็บไซต์ได้ไม่ค่อยดีนัก [3][4] ด้วยเหตุนี้ จึงได้มีการนำเทคนิคการเรียนรู้ของเครื่องมาใช้ในงานตรวจสอบที่อยู่เว็บไซต์อันตราย [5] ซึ่งปัจจัยที่ช่วยส่งเสริมให้แบบจำลองการเรียนรู้ของเครื่องสามารถทำงานได้อย่างมีประสิทธิภาพก็คือ การมีการจัดการข้อมูลที่ดี และ การเลือกคุณลักษณะที่เหมาะสมมาใส่ในแบบจำลอง โดยคุณลักษณะที่สกัดได้จากที่อยู่เว็บไซต์สามารถจำแนกได้เป็น 3 กลุ่มโดยทั่วไป ซึ่งประกอบด้วย คุณลักษณะทางภาษา (Lexical Feature), คุณลักษณะทางเนื้อหา (Content-based feature), และคุณลักษณะทางเครือข่าย (Network-based Feature) [6]

อย่างไรก็ตามผู้วิจัยได้กำหนดขอบเขตในการศึกษาและทดลองในครั้งนี้ โดยเลือกศึกษาและทำการทดลองโดยใช้เพียง 2 กลุ่มคุณลักษณะที่มีความเกี่ยวข้องกับการวิเคราะห์ลักษณะ และโครงสร้างทางภาษา (Textual Properties) เพียง 2 กลุ่มเท่านั้น ซึ่งประกอบไปด้วย กลุ่มคุณลักษณะทางภาษา (Lexical Feature) และ กลุ่มคุณลักษณะทางเนื้อหา (Content-based feature) และใช้เทคนิค Recursive Feature Elimination (RFE) มาช่วยในการคัดเลือกคุณลักษณะจากแต่ละกลุ่ม และ ใช้อัลกอริทึมป่าไม้แบบสุ่ม (Random Forest) ในการสร้างแบบจำลองเพื่อเปรียบเทียบประสิทธิภาพระหว่างกลุ่มคุณลักษณะทั้ง 2 กลุ่ม ในด้านการส่งเสริมความสามารถในการเรียนรู้และความแม่นยำในการทำนายที่อยู่เว็บไซต์ที่เป็นอันตราย ให้แก่แบบจำลองการเรียนรู้ของเครื่อง

1.2 วัตถุประสงค์ของการศึกษา

1.2.1 เพื่อศึกษาและทำความเข้าใจเกี่ยวกับลักษณะของคุณลักษณะทางภาษา (Lexical Feature) และ คุณลักษณะทางเนื้อหา (Content-based Feature)

1.2.2 เพื่อเปรียบเทียบประสิทธิภาพในการส่งเสริมความสามารถในการเรียนรู้ และความแม่นยำในการทำนายที่อยู่เว็บไซต์ให้แก่แบบจำลองการเรียนรู้ของเครื่อง ระหว่างกลุ่มคุณลักษณะทางภาษา (Lexical Feature) และ คุณลักษณะทางเนื้อหา (Content-based Feature)

1.3 ขอบเขตการวิจัย

1.3.1 เลือกศึกษาและทำการทดลอง เพียง 2 กลุ่มคุณลักษณะ คือ กลุ่มคุณลักษณะทางภาษา (Lexical Feature) และ กลุ่มคุณลักษณะทางเนื้อหา (Content-based Feature) เท่านั้น

1.3.2 ใช้เทคนิค Recursive Feature Elimination (RFE) ในขั้นตอนการคัดเลือกคุณลักษณะแต่ละกลุ่มเท่านั้น

1.3.3 ใช้เพียงอัลกอริทึมป่าไม้แบบสุ่ม (Random Forest) ในการทดลองเท่านั้น

1.3.4 ประเมินผลโดยการนำผลการทำนาย (Confusion Matrix) ของแบบจำลองการเรียนรู้ของเครื่อง มาคำนวณหาค่า Accuracy, Precision, recall และ F1 Score และทำการเปรียบเทียบระหว่างแบบจำลองที่ใช้กลุ่มคุณลักษณะทางภาษา (Lexical feature) และ แบบจำลองที่ใช้กลุ่มคุณลักษณะทางเนื้อหา (Content-based feature)

1.3.5 เป็นงานวิจัยที่จัดทำเพื่อศึกษาและเปรียบเทียบผลประสิทธิภาพในการส่งเสริมความแม่นยำในการทำนายที่อยู่เว็บไซต์ให้แก่แบบจำลองการเรียนรู้ ระหว่างกลุ่มคุณลักษณะทางภาษา และ กลุ่มคุณลักษณะทางเนื้อหา ในกรณีที่ใช้วิธีการ รวมถึงเครื่องมือชนิดเดียวกันในการทดลองเท่านั้น

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

ฟิชชิง(Phishing) เป็นภัยคุกคามทางไซเบอร์ที่สร้างความรู้สึกไม่ปลอดภัยและอันตรายแก่ผู้ใช้งานอินเทอร์เน็ต การป้องกันผู้ใช้งานจากที่อยู่เว็บไซต์อันตรายต่างๆ สามารถทำได้หลากหลายวิธีหนึ่งในวิธีที่ถูกแนะนำคือการใช้การเรียนรู้ของเครื่องมาช่วยในการจำแนกที่อยู่เว็บไซต์ต่างๆ เหล่านั้นโดยเนื้อหาที่จะกล่าวถึงในบทนี้จะประกอบไปด้วยเนื้อหาเกี่ยวกับฟิชชิง, องค์ประกอบของที่อยู่เว็บไซต์, การนำเทคนิคการเรียนรู้ของเครื่องมาใช้ในการตรวจสอบที่อยู่เว็บไซต์, คุณลักษณะที่นำมาใช้ในการทดลอง, และงานวิจัยที่เกี่ยวข้อง

2.1 ฟิชชิง (Phishing)

ฟิชชิง คือ หนึ่งในรูปแบบการหลอกลวงทางออนไลน์ โดยผู้โจมตีใช้วิธีการมองหาช่องโหว่จาก พฤติกรรมและลักษณะนิสัยทางธรรมชาติของมนุษย์ เช่น ความอยากรู้อยากเห็น ความตื่นเต้น โดยผู้โจมตีจะสร้างสถานการณ์ที่ทำให้ดูน่าเชื่อถือ จนเหยื่อรู้สึกคล้อยตาม และยอมทำตามสิ่งที่ผู้โจมตีต้องการ โดยเป้าหมายในการโจมตีมักจะเป็น การขโมยเงิน หรือ การขโมยข้อมูลส่วนบุคคล โดยเหยื่อสามารถเป็นได้ทั้งบุคคล และ องค์กร [7] รูปแบบของการฟิชชิงนั้นมีอยู่หลากหลาย เนื่องจากผู้โจมตีมักมีการพัฒนาเทคนิค และกลยุทธ์ในการหลอกลวงเหยื่ออยู่เสมอ ด้วยเหตุนี้จึงทำให้การป้องกันการโจมตีฟิชชิงนั้นเป็นเรื่องยาก [8],[9]

2.1.1 ตัวอย่างการฟิชชิงในรูปแบบต่างๆ

1. การสร้างเว็บไซต์ปลอม (Spoofed Website) คือ รูปแบบการฟิชชิงโดยการสร้างเว็บไซต์ปลอมขึ้นมาโดยเลียนแบบรูปแบบและลักษณะของเว็บไซต์ที่มีอยู่จริง เพื่อสร้างความน่าเชื่อถือ

2. การฟิชชิงผ่าน URL (URL Phishing) คือ การซ่อนลิงก์ โดยผู้โจมตีใช้วิธีหลอกล่อให้เหยื่อกดลิงก์ที่ดูเหมือนที่อยู่เว็บไซต์จริง แต่จริงๆแล้วลิงก์นั้นเชื่อมไปยังเว็บไซต์ปลอมที่แฮกเกอร์สร้างขึ้นเพื่อขโมยข้อมูลของเหยื่อ [10]

3. การโจมตีทางโซเชียลมีเดีย (Social Media Phishing) คือ การโจมตีโดยใช้ช่องทางโซเชียลมีเดียต่างๆ โดยผู้โจมตีอาจใช้วิธีการสร้างบัญชีปลอม แสร้งเป็นบุคคล หรือองค์กรเพื่อเผยแพร่ข้อมูล หรือ ทำการแนบที่อยู่เว็บไซต์ที่เกี่ยวข้องกับการฟิชชิงไปยังผู้ใช้งานในแพลตฟอร์มโซเชียลมีเดียนั้น

4. การโจมตีผ่านเครื่องมือค้นหา (Search Engine Phishing) คือ การสร้างที่อยู่เว็บไซต์ปลอมขึ้นมา โดยเลียนแบบเว็บไซต์ที่มีอยู่จริง เมื่อผู้ใช้งานทำการค้นหาผ่านเครื่องมือค้นหา การแสดงผลลัพธ์ของเว็บไซต์ต่าง ๆ จะรวมถึงเว็บไซต์ปลอมด้วย

5. สเปียร์ฟิชชิง (Spear phishing) คือ รูปแบบการโจมตี ที่เลือกผู้โจมตีจะทำการเลือกเป้าหมายแบบเฉพาะเจาะจง ซึ่งอาจเป็นได้ทั้งบุคคล หรือ องค์กร และทำการศึกษาตลอดจนเก็บข้อมูลเกี่ยวกับเป้าหมายอย่างละเอียดเพื่อนำข้อมูลต่าง ๆ เหล่านั้นมาใช้สร้างสถานการณ์ให้ดูเหมือนเป็นเรื่องจริง และทำให้ดูน่าเชื่อถือมากขึ้น

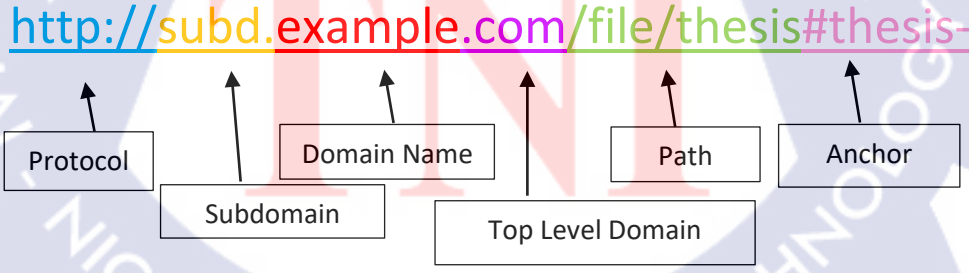
6. การโจมตีโดยการแทรกเนื้อหา (Content Injection Phishing) คือ การโจมตีโดยการแทรกเนื้อหาปลอมหรือลิงก์ที่เกี่ยวข้องกับการฟิชชิง ไปยังเว็บไซต์หรือแอปพลิเคชัน

7. การโจมตีโดยใช้เสียง (Vishing) หรือ เรียกอีกอย่างว่า Voice phishing คือ รูปแบบการโจมตีที่มิจฉาชีพจะปลอมตัวและ ติดต่อทางโทรศัพท์ไปหาเหยื่อ และทำการสร้างสถานการณ์ที่ทำให้ดูน่าเชื่อถือจนทำให้เหยื่อรู้สึกคล้อยตามและเผลอให้ข้อมูลสำคัญไป

8. การโจมตีผ่านข้อความ (Smishing) คือ การขโมยข้อมูลโดยใช้การส่งข้อความ SMS ไปยังเหยื่อ โดยอาจทำการแนบฟิชชิงลิงก์เอาไว้เพื่อให้เหยื่อคลิก และทำการเข้าไปกรอกข้อมูลตามที่ผู้โจมตีต้องการ

2.2 องค์ประกอบของที่อยู่เว็บไซต์

ที่อยู่เว็บไซต์ (URL: Uniform Resource Locator) คือ ตัวระบุที่อยู่ของทรัพยากรต่างๆ บนอินเทอร์เน็ต [11] โดยที่อยู่เว็บไซต์มักประกอบไปด้วย 6 ส่วนดังรูป ที่ 2.1



รูปที่ 2.1 องค์ประกอบของที่อยู่เว็บไซต์

1. โปรโตคอล (Protocol) คือ ส่วนที่อยู่หน้าสุดของที่อยู่เว็บไซต์ มีหน้าที่ในการระบุข้อกำหนดวิธีการในการสื่อสาร หรือเข้าถึงเว็บไซต์ต่าง ๆ โดยโปรโตคอลที่ใช้อยู่โดยทั่วไป เช่น http:// และ https://

2. ซับโดเมน (Subdomain) คือ โดเมนย่อยที่ทำหน้าที่ช่วยนำทางไปยังหน้าต่างๆ ภายในเว็บไซต์ มีความจำเป็นในกรณีเว็บไซต์มีการประกอบไปด้วยส่วนย่อยต่างๆ ภายในเว็บไซต์

3. โดเมนเนม (Domain Name) คือ ชื่อที่ใช้เพื่อระบุและระบุตำแหน่งของเว็บไซต์บนอินเทอร์เน็ต

4. โดเมนระดับบนสุด (Top Level Domain : TLD) คือ ส่วนที่อยู่ต่อท้ายโดเมนเนม โดยสามารถเชื่อมต่อโดยใช้เครื่องหมายจุด (.) มีหน้าที่ คือ ระบุประเภท หรือ ลักษณะของเว็บไซต์ ตัวอย่างของโดเมนเนมระดับบนสุด เช่น .com สำหรับเว็บไซต์ทางธุรกิจ และการค้าขาย, .org สำหรับองค์กรไม่แสดงหากำไร หรือ .co.th สำหรับบริษัทที่ทำการค้าขายในประเทศไทย เป็นต้น

5. พาธ (Path) คือ ส่วนที่นำทางไปสู่ตำแหน่งที่อยู่ของข้อมูล บนเว็บไซต์ โดยจะอยู่หลังโดเมนเนมระดับบนสุดเชื่อมต่อด้วยเครื่องหมายทับ (/)

6. แองเคอร์ (Anchor) จุดที่ปักหมุดไว้บนหน้าเว็บไซต์ ขึ้นต้นด้วยเครื่องหมายชาร์ป (#)

2.3 การเรียนรู้ของเครื่อง (Machine Learning)

การป้องกันการโจมตีแบบฟิชชิง ผ่านทางที่อยู่เว็บไซต์นั้นมีอยู่หลากหลายวิธี [12] การเรียนรู้ของเครื่อง คือ หนึ่งในเทคนิคที่ถูกนำมาใช้ในการตรวจสอบที่อยู่เว็บไซต์ เนื่องจากความสามารถในการเรียนรู้ และความสามารถในการปรับตัวอัตโนมัติตามรูปแบบของข้อมูลจึงทำให้เป็นที่นิยม [13],[14] โดยแบบจำลองการเรียนรู้ของเครื่องจะทำหน้าที่ในการช่วยจำแนกที่อยู่เว็บไซต์ออกเป็นสองกลุ่ม (Binary Classification) คือ กลุ่มที่อยู่เว็บไซต์ที่ปลอดภัย (Benign URL) และอีกกลุ่มคือที่อยู่เว็บไซต์ที่ไม่ปลอดภัย (Malicious URL)

โดย 2 อัลกอริทึมที่เป็นที่นิยมและถูกนำมาใช้ในงานตรวจสอบที่อยู่เป็นเว็บไซต์ คือ Support Vector Machine และ Random Forest นอกจากนี้ก็ยังพบอะกอริทึมอื่นๆอีก เช่น Naïve Bayes, Decision Tree, Logistic Regression และ K-Nearest-Neighbor เป็นต้น [15] อย่างไรก็ตามจากงานวิจัยของ R. Zieni และคณะ [16] ที่ได้ทำการสำรวจงานวิจัยที่เกี่ยวข้องกับการตรวจสอบที่อยู่เว็บไซต์พบว่า งานวิจัยจำนวนมากที่ใช้อัลกอริทึมมากกว่าหนึ่งชุดในการทดลองเพื่อเปรียบเทียบหาอัลกอริทึมที่ดีที่สุดสำหรับการทดลองนั้น ๆ พบว่า แบบจำลองป่าไม้แบบสุ่ม (Random Forest) นั้นมีแนวโน้มที่จะมีประสิทธิภาพที่เหนือกว่าอัลกอริทึมอื่น ๆ ในการทดลองนั้น

2.4 คุณลักษณะ (Features)

คุณลักษณะเป็นหนึ่งในองค์ประกอบที่ช่วยส่งเสริมให้แบบจำลองการเรียนรู้ทำงานได้อย่างมีประสิทธิภาพมากยิ่งขึ้น [17] โดยคุณลักษณะที่ปรากฏในงานวิจัยเกี่ยวกับการตรวจสอบที่อยู่เว็บไซต์นั้นมีอยู่หลากหลายกลุ่ม เช่น คุณลักษณะทางภาษา (Lexical-based Feature), คุณลักษณะทางเนื้อหา (Content-based Feature), คุณลักษณะทางเครือข่าย (Network-based Feature) และ คุณลักษณะที่เกี่ยวข้องกับความนิยม (Popularity Feature) อย่างไรก็ตาม ในงานวิจัยนี้จะทำการศึกษาและทำการทดลองโดยใช้เพียง 2 คุณลักษณะทางภาษา (Lexical Feature) และ คุณลักษณะทางเนื้อหา (Content-based Feature) เท่านั้น ซึ่ง 2 คุณลักษณะดังกล่าวมีความเกี่ยวเนื่องกับการวิเคราะห์เชิงภาษา โดยคุณลักษณะทางภาษาเกี่ยวข้องกับ โครงสร้าง และ องค์ประกอบ ของตัวที่อยู่เว็บไซต์นั้น ๆ ในขณะที่ลักษณะทางเนื้อหา จะเกี่ยวข้องไปถึงหน้าเนื้อหาที่เชื่อมโยงกับเว็บไซต์นั้น ๆ ซึ่งหมายถึงสิ่งที่ปรากฏบน HTML code หรือ JavaScript

2.4.1 คุณลักษณะทางภาษา (Lexical Feature)

คุณลักษณะในกลุ่มทางภาษา จะมีความเกี่ยวข้องกับสิ่งที่ปรากฏบนที่อยู่เว็บไซต์ เช่น ความยาวของที่อยู่เว็บไซต์ ความยาวชื่อโดเมน จำนวนของตัวอักษร ตัวเลข รวมถึงอักขระพิเศษต่าง ๆ [18] โดยตัวอย่างของคุณลักษณะในกลุ่มคุณลักษณะทางภาษาปรากฏในตารางที่ 2.1

ตารางที่ 2.1 ตัวอย่างคุณลักษณะทางภาษา [18]

No.	Feature group	Feature	Data type	Description
1	Lexical group	SymbolCount_URL	numeric	Number of symbol in URL
2		urlLen	numeric	The length of URL
3		domainlength	numeric	The length of domain
4		URL_Letter_Count	numeric	Number of the Letter
5		URL_DigitCount	numeric	Number of the numeric Character

2.4.2 คุณลักษณะทางเนื้อหา (Content-based Feature)

คุณลักษณะทางเนื้อหา คือคุณลักษณะที่มาจากการวิเคราะห์สิ่งที่ปรากฏบนหน้าเว็บไซต์ที่เชื่อมต่อกับที่อยู่เว็บไซต์นั้น ๆ โดยวิเคราะห์จาก HTML code หรือ JavaScript เช่น จำนวนของแท็ก HTML หรือ องค์ประกอบต่างๆ เช่น จำนวน ลิงก์เชื่อมโยง (Hyperlinks) เป็นต้น โดยตารางที่ 2.2 เป็นตัวอย่างของคุณลักษณะในกลุ่มคุณลักษณะทางเนื้อหา

ตารางที่ 2.2 ตัวอย่างคุณลักษณะทางเนื้อหา [19]

No.	Feature group	Feature	Description
1	Content-based group	Length of HTML	Total number of characters in HTML pages excluding tags
2		Number Script Tags	Total number of scripts included in the page
3		Number iframe	Total number of iframe tags on page
4		Number of internal hyperlinks	Total number of internal hyperlinks on page
5		Number of external hyperlinks	Total number of external hyperlinks on page

2.5 งานวิจัยที่เกี่ยวข้อง

จากการศึกษางานวิจัยในกลุ่มที่ศึกษาเกี่ยวกับการจำแนกที่อยู่เว็บไซต์นั้น พบว่างานวิจัยส่วนใหญ่มุ่งเน้นไปที่การเปรียบเทียบประสิทธิภาพของ อัลกอริทึม โดยเลือกใช้อัลกอริทึมที่หลากหลายมาทำการทดลอง อย่างไรก็ตามพบว่า หนึ่งในปัจจัยที่ช่วยส่งเสริมความแม่นยำและความเร็วในการจำแนกที่อยู่เว็บไซต์ต่าง ๆ นั้น คือ คุณลักษณะ (Features) [20], โดยคุณลักษณะที่ปรากฏในงานวิจัยจำแนกที่อยู่เว็บไซต์อันตรายนั้น มีอยู่หลากหลายกลุ่ม เช่น กลุ่มคุณลักษณะทางภาษา (Lexical Feature) คุณลักษณะที่เกี่ยวข้องกับโฮสต์ (Host-based Feature), และ คุณลักษณะที่มีความสัมพันธ์กัน (Correlated feature) เป็นต้น โดยกลุ่มคุณลักษณะที่พบบ่อย ๆ จะมีอยู่ 3 กลุ่ม คือ กลุ่มคุณลักษณะทางภาษา (Lexical Feature), กลุ่มคุณลักษณะทางเนื้อหา (Content-based Feature), และ กลุ่มคุณลักษณะทางเครือข่าย (Network-based Feature)

โดยงานวิจัยในกลุ่มแรกที่ผู้วิจัยศึกษา คือ งานวิจัยในกลุ่มที่ศึกษาโดยใช้กลุ่มคุณลักษณะทางภาษาในการทดลอง โดยงานวิจัย M. Elsadig และคณะ [21] ทำการศึกษาเกี่ยวกับการจำแนกที่อยู่เว็บไซต์โดยใช้ โมเดล BERT (Bidirectional Encoder Representations from Transformers) ในขั้นตอนการสกัดคุณลักษณะของที่อยู่เว็บไซต์เพื่อช่วยวิเคราะห์โครงสร้างทางภาษาของที่อยู่เว็บไซต์ จำนวน 549,346 รายการในชุดข้อมูลที่นำมาใช้ในการทดลอง โดย Deep Convolutional Neural Network (CNN) มีผลลัพธ์ความแม่นยำสูงถึง 96.66% งานวิจัยของ A. S. Raja และคณะ [22] เป็นงานวิจัยที่ศึกษาโดยมีจุดมุ่งหมายเพื่อหาวิธีในการช่วยลดเวลาในการประมวลผล และลดพื้นที่สำหรับการจัดเก็บ ได้ทำการศึกษา โดยใช้เพียงคุณลักษณะทางภาษา (Lexical Feature) จำนวน 20 คุณลักษณะ โดยใช้เทคนิค Correlation analysis มาช่วยในการลดจำนวนคุณลักษณะจากเดิมมีทั้งสิ้น 27 คุณลักษณะ มาทำการทดลองโดยใช้แบบจำลองการเรียนรู้ของเครื่องที่แตกต่างกัน โดยแบบจำลองป่าไม้แบบสุ่ม (Random Forest) ได้ค่าความแม่นยำอยู่ที่ 99.99% นอกจากนี้ ยังมีงานวิจัยที่ศึกษาโดย W.Faiq and N.G.M. Jameel [23] นำเสนอแบบจำลอง Multilayer perceptron (MLP) ซึ่งเป็นอีกหนึ่งวิธีที่สามารถช่วยลดเวลาในการประมวลผลได้ โดยในการทดลองนี้ใช้คุณลักษณะทางภาษาจำนวน 35 คุณลักษณะ โดยแบบจำลองดังกล่าวมีความแม่นยำอยู่ที่ 94.51%. อีกงานวิจัยที่ศึกษาโดย B.Atlay et al. [24] คณะผู้วิจัยได้ทำการศึกษาและทดลองโดยใช้ชุดข้อมูลที่ถูกเก็บรวบรวมจากเว็บไซต์ PhishTank และ Alexa โดยคุณลักษณะทางเนื้อหา (Content-based Feature) ถูกสกัดโดยใช้ keyword density extractor library ที่ถูกพัฒนาโดย Comodo Group. ซึ่งค่าความแม่นยำสูงสุดอยู่ที่ 98.24% โดยแบบจำลอง Support Vector Machine (SVM

ถัดมาจะเป็นงานวิจัยที่ศึกษาและทดลองโดยใช้คุณลักษณะในการทดลองมากกว่าหนึ่งกลุ่มคุณลักษณะขึ้นไป งานวิจัยศึกษาโดย S. Kumi และคณะ [25] นำเสนอวิธีการจำแนกที่อยู่เว็บไซต์โดยใช้ เทคนิค Classification based on association (CBA) โดยใช้กลุ่มคุณลักษณะทางภาษา และกลุ่มคุณลักษณะทางเนื้อหา โดยได้ความแม่นยำในการทำนายสูงสุดอยู่ที่ 95.8% ต่อมาเป็นงานวิจัยที่ศึกษาโดย J. Liu และคณะ [26] ทำการทดลองโดยใช้คุณลักษณะทางภาษา และ คุณลักษณะทางเนื้อหา แต่ในงานวิจัยนี้นั้น ผู้วิจัยได้ทำการสกัดคุณลักษณะจาก โค้ดต้นฉบับหน้าหลัก (Homepage source code) หลังจากนั้นนำมาทดลองโดยใช้แบบจำลองที่หลากหลาย แต่แบบจำลองที่ให้ค่าความแม่นยำสูงที่สุด คือ แบบจำลองป่าไม้สุ่ม (Random forest) โดยค่าความแม่นยำอยู่ที่ 93% ต่อมาเป็นงานวิจัยที่ศึกษาโดย P.V. Rao และคณะ [27] ทำการทดลองโดยใช้ honeypot และ PyShark ในการเก็บข้อมูลเกี่ยวกับการโต้ตอบกับที่อยู่เว็บไซต์ต่างๆ และการจราจรผ่านเครือข่าย โดยคุณลักษณะที่ถูกใช้ในงานนี้ประกอบไปด้วย คุณลักษณะทางภาษา, คุณลักษณะทางเนื้อหา, และ คุณลักษณะทางเครือข่าย โดยแบบจำลอง XGBoost ได้ค่าความแม่นยำสูงสุด โดยมีค่าความแม่นยำอยู่ที่ 96.8%

เพื่อแก้ปัญหาเมื่อเจอกับคุณลักษณะที่มีจำนวนเยอะ ๆ นั้น M. N. Alam และคณะ [28] ได้ใช้เทคนิค Principal Component Analysis (PCA) มาช่วยในการลดมิติของข้อมูล เพื่อให้เห็นความสัมพันธ์ของข้อมูลง่ายขึ้น และช่วยในขั้นตอนการคัดเลือกคุณลักษณะ ซึ่งผลการทดลองของงานวิจัยนี้นั้น มีค่าความแม่นยำอยู่ที่ 97% ซึ่งเป็นผลการทำนายที่อยู่เว็บไซต์โดยแบบจำลองป่าไม้แบบสุ่ม (Random forest) อีกงานวิจัยที่ศึกษาโดย M. Aljabi และคณะ [29] เลือกใช้ 3 เทคนิค ในขั้นตอนการคัดเลือกคุณลักษณะ เพื่อช่วยคัดเลือกคุณลักษณะจาก 3 กลุ่ม คุณลักษณะ คือ กลุ่มคุณลักษณะทางเนื้อหา, กลุ่มคุณลักษณะทางภาษา, และ กลุ่มคุณลักษณะทางเครือข่าย ซึ่ง 3 เทคนิคที่ถูกนำมาใช้ในขั้นตอนการคัดเลือกคุณลักษณะนั้น ประกอบไปด้วย เทคนิค ANOVA, เทคนิค Chi-square, และเทคนิค Correlation โดยคัดเลือกคุณลักษณะจำนวนทั้งหมด 40 คุณลักษณะ (รวมทั้ง 3 กลุ่มคุณลักษณะ) เหลือเพียง 15 คุณลักษณะ ซึ่งผลการทดลองออกมาว่า แบบจำลอง Naïve Bayes นั้นได้ค่าความแม่นยำสูงที่สุดเมื่อเทียบกับแบบจำลองอื่น ๆ ที่ถูกนำมาทดลองด้วย โดยความแม่นยำในการทำนายที่อยู่เว็บไซต์อันตรายนั้นอยู่ที่ 96%

และในงานวิจัยสุดท้ายที่ถูกยกมากล่าวถึงในงานวิจัยนั้น เป็นการศึกษาของ C. D. Xuan และคณะ [30] โดยเน้นศึกษาไปที่การเปรียบเทียบ ปริมาณของข้อมูลที่ใช้ในการสอนแบบจำลอง โดยแบ่งการทดลองออกเป็น 2 รอบ โดยรอบแรก ใช้ข้อมูลที่อยู่เว็บไซต์จำนวน 10,000 รายการ และการทดลองที่ 2 อยู่ที่ 470,000 รายการ โดยคุณลักษณะที่เลือกใช้นั้น ประกอบไปด้วย คุณลักษณะทางภาษา, คุณลักษณะที่เกี่ยวข้องกับโฮสต์ (Host-based feature), และ คุณลักษณะที่มีความสัมพันธ์กัน (Correlated feature) โดย ผลการทดลองนั้น พบว่า การทดลองด้วย 10,000 ที่อยู่เว็บไซต์มีความแม่นยำสูงกว่า โดยแบบจำลอง

ป่าไม้แบบสุ่ม (Random Forest) มีค่าความแม่นยำอยู่ที่ 99.77% อย่างไรก็ตามเพื่อให้เห็นภาพรวมของงานวิจัยต่าง ๆ ผู้วิจัยได้ทำการสรุปงานวิจัยที่เกี่ยวข้องอีกครั้ง ตามตารางที่ 2.3

ตารางที่ 2.3 ตารางสรุปงานวิจัยที่เกี่ยวข้อง

ชื่องานวิจัย	กลุ่มคุณลักษณะ	วิธีการ	ค่าความแม่นยำ
Intelligent Deep Machine Learning Cyber Phishing URL Detection Based on Bert Features Extraction [21]	คุณลักษณะทางภาษา	Deep Convolutional Neural Network (CNN)	96.66%
Lexical features based malicious URL detection using machine learning techniques [22]	คุณลักษณะทางภาษา	Random Forest	99.99%
Malicious URL Detection Tree-based Lexical Features Selection and Multilayer Perceptron model [23]	คุณลักษณะทางภาษา	Multilayer perceptron (MLP)	94.51%
Context-sensitive and keyword density based supervised machine learning techmiques for malicious webpage detection [24]	คุณลักษณะทางเนื้อหา	SVM (Support Vector Machine)	98.24%

ตารางที่ 2.3 ตารางสรุปงานวิจัยที่เกี่ยวข้อง (ต่อ)

ชื่องานวิจัย	กลุ่มคุณลักษณะ	วิธีการ	ค่าความแม่นยำ
Malicious URL Detection Based on Associative Classification [25]	คุณลักษณะทางภาษา และ คุณลักษณะทางเนื้อหา	Classification based on association (CBA)	95.8%
Detecting Web Spam Based on Novel Features From Web Page Source Code [26]	คุณลักษณะทางภาษา และ คุณลักษณะทางเนื้อหา	Random Forest	93%
Detection of Malicious Uniform Resource Locator [27]	คุณลักษณะทางภาษา, คุณลักษณะทางเนื้อหา และ คุณลักษณะทางเครือข่าย	XGBoost	96.8%
Phishing Attacks Detection using Machine Learning Approach [28]	คุณลักษณะทางภาษา, คุณลักษณะทางเนื้อหา และ คุณลักษณะทางเครือข่าย	Random forest	97%
An Assessment of Lexical, Network, and Content-Based Features for Detecting Malicious URLs Using Machine Learning and Deep Learning Models [29]	คุณลักษณะทางภาษา, คุณลักษณะทางเนื้อหา และ คุณลักษณะทางเครือข่าย	Random forest	97%
Malicious URL Detection based on Machine Learning [30]	คุณลักษณะทางภาษา, คุณลักษณะที่เกี่ยวข้องกับโฮสต์, และ คุณลักษณะที่มีความสัมพันธ์กัน	Random forest	99.77%

บทที่ 3
วิธีการดำเนินงาน

งานวิจัยนี้เป็นการศึกษาเกี่ยวกับการจำแนกที่อยู่เว็บไซต์ออกเป็นสองกลุ่ม คือ กลุ่มที่อยู่เว็บไซต์ที่ไม่ปลอดภัย (Malicious URL) และ กลุ่มที่อยู่เว็บไซต์ที่ปลอดภัย (Benign URL) ด้วยวิธีการเรียนรู้ของเครื่อง โดยมีจุดประสงค์ในการศึกษา และ ทำการทดลอง เพื่อหาคุณลักษณะที่ช่วยส่งเสริมความแม่นยำให้กับแบบจำลองการเรียนรู้ของเครื่องในงานจำแนกที่อยู่เว็บไซต์ได้ดีที่สุด โดยงานวิจัยนี้ผู้วิจัยเลือกกลุ่มคุณลักษณะ เพียง 2 กลุ่ม ที่สามารถสกัดได้จากที่อยู่เว็บไซต์ คือ กลุ่มคุณลักษณะทางภาษา (Lexical Features) และ กลุ่มคุณลักษณะทางเนื้อหา (Content-based Features) มาทำการศึกษาและทดลองเปรียบเทียบประสิทธิภาพเท่านั้น โดยเนื้อหาในบทนี้จะเป็นการอธิบายในส่วนของขั้นตอนในการดำเนินงานวิจัย ซึ่งสามารถแบ่งออกได้เป็น 4 ขั้นตอนตามรูปภาพที่ 3.1 คือ

1. การเก็บข้อมูล (Data Collection)
2. การเตรียมข้อมูล (Data Preprocessing)
3. การสร้างโมเดล (Modeling)
4. การประเมินผล (Evaluation)

รูปที่ 3.1 ขั้นตอนในการดำเนินงานวิจัย

3.1 การจัดเก็บข้อมูล (Data Collection)

ชุดข้อมูลที่นำมาใช้ในการทดลองครั้งนี้เป็นชุดข้อมูลออนไลน์ “Malicious and Benign Webpages Dataset” โดย A.K. Singh [31] ข้อมูลถูกจัดเก็บโดยใช้ MalCrawler ซึ่งเป็นเครื่องมือสำหรับการตรวจหา และดึงข้อมูลเว็บไซต์ที่ไม่ปลอดภัย (Malicious Websites) จากอินเทอร์เน็ต และใช้ Google Safe Browsing API ในการกำหนดว่าที่อยู่เว็บไซต์ดังกล่าวปลอดภัยหรือไม่ปลอดภัย ซึ่งในชุดข้อมูลนี้ประกอบไปด้วย 2 ชุดข้อมูล คือ ชุดข้อมูลสำหรับการสอน (Train dataset) และ ชุดข้อมูลสำหรับการทดสอบ (Test Dataset) โดยชุดข้อมูลสำหรับการสอนนั้นบรรจุที่อยู่เว็บไซต์จำนวน 1,200,000 รายชื่อ โดยแบ่งเป็นรายชื่อเว็บไซต์ที่ปลอดภัย 1,172,747 รายชื่อ และ รายชื่อเว็บไซต์ที่ไม่ปลอดภัย 27,253 รายชื่อ ในขณะที่ชุดข้อมูลทำหรับการทดสอบ บรรจุรายชื่อเว็บไซต์ที่ปลอดภัย 353,872 รายชื่อ และ รายชื่อเว็บไซต์ที่ไม่ปลอดภัยจำนวน 8,062 รายชื่อ รวมรายชื่อที่อยู่เว็บไซต์ในชุดข้อมูลสำหรับการทดสอบทั้งหมด 361,934 รายชื่อ นอกจากนี้ในชุดข้อมูลที่อยู่เว็บไซต์ดังกล่าวนี้ยังประกอบไปด้วยคุณลักษณะของข้อมูลอีก 11 รายการ ตามปรากฏในตารางที่ 3.1

ตาราง 3.1 คุณลักษณะที่ปรากฏในชุดข้อมูล

ชื่อคุณลักษณะ	คำอธิบาย
1. url	URL ของเว็บไซต์
1. ip_add	ที่อยู่ IP ของเว็บไซต์
2. geo_loc	ตำแหน่งทางภูมิศาสตร์ของเว็บไซต์
3. url_len	ความยาวของ URL
4. js_len	ความยาวของสคริปต์ JavaScript ในหน้าเว็บ
5. js_obf_len	ความยาวของสคริปต์ JavaScript หลังจากการบดบัง
6. tld	คือส่วนขยายท้ายของ URL เช่น .com, หรือ .org เป็นต้น
7. who_is	ข้อมูลโดเมนที่ได้มาจากการใช้งาน WHO IS นั้นครบถ้วนหรือไม่
8. https	เป็นเว็บไซต์ที่ใช้โปรโตคอล HTTPS หรือไม่
9. content	เนื้อหาของหน้าเว็บไซต์ต้นฉบับรวมถึงโค้ด JavaScript
10. label	ป้ายกำกับว่าเว็บไซต์ดังกล่าวเป็นเว็บไซต์ปลอดภัยหรือไม่ปลอดภัย

3.2 การเตรียมข้อมูล (Data preprocessing)

ขั้นตอนการเตรียมข้อมูลนั้นถือเป็นขั้นตอนพื้นฐานที่สำคัญเพื่อเตรียมข้อมูลให้พร้อมสำหรับการนำไปใช้งานการศึกษาของแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning) โดยในขั้นตอนนี้สามารถแบ่งย่อยได้เป็น 3 ขั้นตอนดังนี้

3.2.1 การลดจำนวนข้อมูล (Undersampling)

จากการวิเคราะห์ชุดข้อมูลที่นำมาใช้ในการทดลองครั้งนี้ พบว่าจำนวนของที่อยู่เว็บไซต์ปลอดภัย และเว็บไซต์ที่ไม่ปลอดภัยนั้นมีความไม่สมดุลกัน เพื่อแก้ปัญหาดังกล่าว ผู้วิจัยจึงได้เลือกใช้วิธีการสุ่มลดจำนวนข้อมูล (Random Undersampling) ซึ่งเป็นหนึ่งในวิธีการที่ใช้สำหรับแก้ปัญหาความไม่สมดุลกันของข้อมูล โดยทำการลดจำนวนข้อมูลในคลาสที่มีจำนวนมากให้มีจำนวนข้อมูลสมดุลกันกับคลาสที่มีจำนวนน้อยกว่า โดยผลจากการใช้เทคนิคดังกล่าวทำให้จำนวนทั้งหมดของที่อยู่เว็บไซต์ที่อยู่ในชุดข้อมูลสำหรับการสอน (Train data) จาก 1,200,000 ที่อยู่เว็บไซต์ เหลือเพียง 54,506 ที่อยู่เว็บไซต์ โดยแบ่งเป็นที่อยู่เว็บไซต์อันตราย (Malicious URL) 27,253 ที่อยู่เว็บไซต์ และ ที่อยู่เว็บไซต์ที่ปลอดภัย (Benign URL) 27,253 ที่อยู่เว็บไซต์ ส่วนจำนวนข้อมูลในชุดข้อมูลสำหรับการทดสอบ (Test data) จากเดิมมีจำนวนที่อยู่เว็บไซต์ทั้งหมด 361,934 ที่อยู่เว็บไซต์ เหลือเพียง 16,124 ที่อยู่เว็บไซต์ โดยแบ่งเป็นที่อยู่เว็บไซต์อันตราย (Malicious URL) จำนวน 8,062 ที่อยู่เว็บไซต์ และ เว็บไซต์ที่ปลอดภัย (Benign URL) จำนวน 8,062 ที่อยู่เว็บไซต์

3.2.2 การสกัดคุณลักษณะของข้อมูล (Feature Extraction)

เนื่องจากคุณลักษณะของข้อมูลทั้ง 11 รายการ ตามตารางที่ 3.1 ที่ปรากฏในชุดข้อมูลนั้น ยังไม่เพียงพอและครบตามกลุ่มคุณลักษณะที่ต้องการสำหรับการทดลองในงานวิจัยนี้ ซึ่งต้องการเปรียบเทียบประสิทธิภาพของคุณลักษณะระหว่างกลุ่มคุณลักษณะทางภาษา (Lexical Feature) และ กลุ่มคุณลักษณะทางเนื้อหา (Content-based Feature) ดังนั้นจึงต้องมีการสกัดคุณลักษณะเพิ่มเติม โดยคุณลักษณะทางภาษานั้น คือ ลักษณะทางภาษาที่ปรากฏอยู่ในที่อยู่เว็บไซต์ต่างๆ เช่น ความยาวของที่อยู่เว็บไซต์, จำนวนอักขระพิเศษต่างๆ เป็นต้น โดยตารางที่ 3.2 ประกอบไปด้วยคุณลักษณะทางภาษาที่สกัดได้ และ คำอธิบายของคุณลักษณะ

ตารางที่ 3.2 คุณลักษณะทางภาษาที่ถูกสกัดจากชุดข้อมูล

ชื่อคุณลักษณะ	คำอธิบาย
1. url_length	ความยาวของที่อยู่เว็บไซต์
2. url_path_length	จำนวนตัวอักษรทั้งหมดในส่วนของเส้นทาง URL
3. url_host_length	จำนวนตัวอักษรทั้งหมดในส่วนของโฮสต์หรือโดเมน
4. url_slash_count	จำนวนเครื่องหมาย /
5. url_dot_count	จำนวนเครื่องหมาย .
6. url_ampersand_count	จำนวนเครื่องหมาย &
7. url_at_symbol_count	จำนวนเครื่องหมาย @
8. url_hyphen_count	จำนวนเครื่องหมาย -
9. url_equals_sign_count	จำนวนเครื่องหมาย =
10. url_question_mark_count	จำนวนเครื่องหมาย ?
11. url_semicolon_count	จำนวนเครื่องหมาย ;
12. number_of_digits	จำนวนตัวเลขที่ปรากฏบนที่อยู่เว็บไซต์
13. number_of_parameters	จำนวนของพารามิเตอร์ที่ปรากฏใน URL
14. number_of_fragments	จำนวนของ fragment identifier เมื่อมีการใช้ #
15. number_of_subdirectories	จำนวนของโฟลเดอร์ย่อยที่ปรากฏในเส้นทาง
16. has_client_in_string_encoded	เช็คคำว่า “client” มี/ไม่มี
17. has_admin_in_string_encoded	เช็คคำว่า “admin” มี/ไม่มี
18. has_server_in_string_encoded	เช็คคำว่า “server” มี/ไม่มี
19. has_login_in_string_encoded	เช็คคำว่า “log_in” มี/ไม่มี

ซึ่งจากตารางที่ 3.2 จะพบว่าบางคุณลักษณะ เช่น คุณลักษณะในลำดับที่ 16 ถึง 19 มีความจำเป็นต้องแปลงข้อมูลจากตัวอักษรเป็นตัวเลข ด้วยเทคนิคการเข้ารหัสข้อมูล (Encoding) โดยกำหนดให้หากผลลัพธ์จากการสกัดแสดงผลว่า มี (True) เท่ากับ 1 ไม่มี (False) เท่ากับ 0 ก่อนนำไปใช้กับแบบจำลองการเรียนรู้ของเครื่อง

ถัดมาในส่วนของคุณลักษณะทางเนื้อหา (Content-based Features) นั้นเป็นอีกหนึ่งคุณลักษณะที่มีความเกี่ยวข้องกับการวิเคราะห์ลักษณะของโครงสร้างทางภาษา (Textual properties) เช่นเดียวกับกลุ่มคุณลักษณะทางภาษา แต่มีความแตกต่างกันตรงที่ที่คุณลักษณะทางเนื้อหานั้นจะไม่ได้วิเคราะห์จากโครงสร้างหรือองค์ประกอบที่ปรากฏบนชื่อของที่อยู่เว็บไซต์ แต่จะเกี่ยวข้องกับการวิเคราะห์ลักษณะทางภาษาที่ปรากฏในส่วนของ JavaScript หรือ HTML code ซึ่งเป็นส่วนของการแสดงเนื้อหาภายหน้าเว็บไซต์ที่เชื่อมต่อกับที่อยู่เว็บไซต์นั้น ๆ โดยตารางที่ 3.3 แสดงผลคุณลักษณะทางเนื้อหาที่สกัดได้ และ คำอธิบายของแต่ละคุณลักษณะ

ตารางที่ 3.3 คุณลักษณะทางเนื้อหาที่ถูกสกัดจากชุดข้อมูล

ชื่อคุณลักษณะ	คำอธิบาย
1. js_len	ความยาวของส่วนโค้ด JavaScript
2. js_obf_len	ความยาวของสคริปต์ JavaScript หลังจากการบดบัง
3. url_page_entropy	ค่าเอ็นโทรปี ของ URL ของหน้าเว็บไซต์
4. number_of_script_tags	จำนวนแท็กสคริปต์ในหน้าเว็บไซต์
5. script_to_body_ratio	อัตราส่วนของสคริปต์ต่อส่วนกลางของหน้าเว็บไซต์
6. length_of_html	ความยาวของโค้ด HTML ในหน้าเว็บไซต์
7. number_of_page_tokens	จำนวนโทเคน (Tokens) ในหน้าเว็บไซต์
8. number_of_sentences	จำนวนประโยคในหน้าเว็บไซต์
9. number_of_punctuations	จำนวนเครื่องหมายวรรคตอนในหน้าเว็บไซต์
10. number_of_distinct_tokens	จำนวนโทเคนที่แตกต่างกันในหน้าเว็บไซต์
11. number_of_capitalizations	จำนวนการใช้ตัวอักษรใหญ่ในหน้าเว็บไซต์

ตารางที่ 3.3 คุณลักษณะทางเนื้อหาที่ถูกสกัดจากชุดข้อมูล (ต่อ)

ชื่อคุณลักษณะ	**คำอธิบาย**
12. average_number_of_tokens_in_sentence	จำนวนโทเคนเฉลี่ยในแต่ละประโยคของหน้าเว็บไซต์
13. number_of_html_tags	จำนวนแท็ก HTML ในหน้าเว็บไซต์
14. number_of_hidden_tags	จำนวนแท็กที่ถูกซ่อนอยู่ในหน้าเว็บไซต์
15. number_iframes	จำนวนของแท็ก iframe ในหน้าเว็บไซต์
16. number_objects	จำนวนวัตถุในหน้าเว็บไซต์ เช่น รูปภาพ
17. number_embeds	จำนวนแท็ก embed ในหน้าเว็บไซต์
18. number_of_hyperlinks	จำนวน Hyperlink ในหน้าเว็บไซต์
19. number_of_whitespace	จำนวนช่องว่างในหน้าเว็บไซต์
20. number_of_included_elements	จำนวนองค์ประกอบต่างๆ ที่ถูกรวมเข้ามาในหน้าเว็บไซต์
21. number_of_double_documents	จำนวนเอกสารที่ซ้ำซ้อนในหน้าเว็บไซต์
22. number_of_eval_functions	จำนวนการใช้ฟังก์ชัน eval ในหน้าเว็บไซต์
23. average_script_length	ความยาวของสคริปต์เฉลี่ยในหน้าเว็บไซต์
24. average_script_entropy	เอ็นโทรปีเฉลี่ยของสคริปต์ในหน้าเว็บไซต์

3.2.3 การคัดเลือกคุณลักษณะ (Feature Selection)

การคัดเลือกคุณสมบัติเป็นอีกหนึ่งวิธีการที่ช่วยเพิ่มความแม่นยำให้กับ การเรียนรู้ของเครื่อง [32] โดยการลดขนาดของคุณลักษณะที่มีอยู่ให้อยู่ในระดับที่เหมาะสมเท่าที่จำเป็น เพื่อลดความซับซ้อนของโมเดล และลดโอกาสในการเกิด Overfitting ซึ่งเป็นปัญหาที่อาจเกิดขึ้นเมื่อโมเดลมีคุณลักษณะมากเกินไปหรือมีคุณลักษณะที่ไม่สำคัญซึ่งอาจทำให้ผลลัพธ์ที่ได้ไม่แม่นยำ ในการทดลองนี้ ผู้วิจัยได้เลือกใช้เทคนิค Recursive Feature Elimination (RFE) มาช่วยในกระบวนการคัดเลือกคุณลักษณะที่สำคัญสำหรับสร้างโมเดลในการจำแนกที่อยู่เว็บไซต์อันตราย (Malicious URL)

โดย RFE เป็นกระบวนการที่ใช้ในการลดจำนวนคุณลักษณะแบบอัตโนมัติโดยการทำซ้ำ โดยเริ่มต้นด้วยการสร้างโมเดล จากนั้นคำนวณค่าความสำคัญของคุณลักษณะแต่ละตัว และเลือกคุณลักษณะที่มีค่าความสำคัญต่ำที่สุดออกไป ต่อมาจะทำซ้ำกระบวนการนี้โดยลบคุณลักษณะที่เหลือออกไปตามจำนวนที่กำหนด จนกว่าจะได้จำนวนคุณลักษณะที่ต้องการ ซึ่งผลลัพธ์ที่ได้จะเป็นลำดับความสำคัญของคุณลักษณะ ซึ่งจะเรียงลำดับจากคุณลักษณะที่มีความสำคัญมากที่สุด โดยตารางที่ 3.4 แสดงคุณลักษณะที่สำคัญ 10 อันดับแรกจากกลุ่มคุณลักษณะทางภาษา และตารางที่ 3.5 แสดงคุณลักษณะที่สำคัญ 10 อันดับแรกจากกลุ่มคุณลักษณะทางเนื้อหา

ตารางที่ 3.4 คุณลักษณะที่สำคัญ 10 อันดับแรก ของกลุ่มคุณลักษณะทางภาษา

ชื่อคุณลักษณะ	คำอธิบาย
1. url_length	ความยาวของที่อยู่เว็บไซต์
2. url_path_length	จำนวนตัวอักษรทั้งหมดในส่วนของเส้นทาง URL
3. url_host_length	จำนวนตัวอักษรทั้งหมดในส่วนของโฮสต์หรือโดเมน
4. url_slash_count	จำนวนเครื่องหมาย /
5. url_dot_count	จำนวนเครื่องหมาย .
6. url_hyphen_count	จำนวนเครื่องหมาย -
7. url_question_mark_count	จำนวนเครื่องหมาย ?
8. number_of_digit	จำนวนตัวเลขที่ปรากฏบนที่อยู่เว็บไซต์
9. number_of_parameter	จำนวนของพารามิเตอร์ที่ปรากฏใน URL
10. number_of_subdirectories	จำนวนของโฟลเดอร์ย่อยที่ปรากฏในเส้นทาง

ตารางที่ 3.5 คุณลักษณะสำคัญ 10 อันดับแรก ของกลุ่มคุณลักษณะทางเนื้อหา

ชื่อคุณลักษณะ	คำอธิบาย
1. js_len	ความยาวของส่วนโค้ด JavaScript
2. js_obf_len	ความยาวของสคริปต์ JavaScript หลังจากการบดบัง
3. script_to_body_ratio	อัตราส่วนของสคริปต์ต่อส่วนกลางของหน้าเว็บไซต์
4. length_of_html	ความยาวของโค้ด HTML ในหน้าเว็บไซต์
5. number_of_punctuations	จำนวนเครื่องหมายวรรคตอนในหน้าเว็บไซต์
6. number_of_distinct_tokens	จำนวนโทเคนที่แตกต่างกันในหน้าเว็บไซต์
7. average_number_of_tokens_in_sentence	จำนวนโทเคนเฉลี่ยในแต่ละประโยคของหน้าเว็บไซต์
8. number_of_html_tags	จำนวนแท็ก HTML ในหน้าเว็บไซต์
9. number_of_hyperlinks	จำนวน Hyperlink ในหน้าเว็บไซต์
10. number_of_included_elements	จำนวนองค์ประกอบต่างๆ ที่ถูกรวมเข้ามาในหน้าเว็บไซต์

3.3 การสร้างโมเดล (Modeling)

วิธีการเรียนรู้ของเครื่อง (Machine learning) เป็นหนึ่งในวิธีที่เป็นที่นิยมนำมาใช้ในการจำแนกที่อยู่เว็บไซต์ที่ไม่ปลอดภัย (Malicious URL) และที่อยู่เว็บไซต์ที่ปลอดภัย (Benign URL) โดยจากงานวิจัยเชิงสำรวจของ Z. Rasha et al. [16] พบว่า งานวิจัยที่ทำการศึกษาเกี่ยวกับการจำแนกที่อยู่เว็บไซต์ ที่ได้ทำการทดลองโดยเปรียบเทียบหลาย ๆ อัลกอริทึมนั้น อัลกอริทึมที่พบได้บ่อย และ มีแนวโน้มว่ามีประสิทธิภาพในการจำแนกที่อยู่เว็บไซต์ได้ดีที่สุดนั้น คือ อัลกอริทึมป่าไม้แบบสุ่ม (Random Forest)

ซึ่งในงานวิจัยนี้ ผู้วิจัยได้ทำการศึกษา และทดลองโดยใช้ อัลกอริทึมป่าไม้แบบสุ่ม (Random forest) โดยการสร้างโมเดล 2 โมเดลสำหรับการตรวจสอบที่อยู่เว็บไซต์อันตราย (Malicious URL) เพื่อทำการเปรียบเทียบประสิทธิภาพระหว่างกลุ่มคุณลักษณะ 2 กลุ่ม โดยโมเดลที่ 1 จะใช้ คุณลักษณะสำคัญ 10 อันดับแรกจากกลุ่มคุณลักษณะทางภาษา (Lexical feature) และโมเดล ที่ 2 ใช้คุณลักษณะสำคัญ 10 อันดับแรกจากกลุ่มคุณลักษณะทางเนื้อหา (Content-based feature)

3.4 การประเมินผล (Evaluation)

สำหรับการประเมินผลการทดลองเพื่อเปรียบเทียบประสิทธิภาพของคุณลักษณะทั้ง 2 กลุ่ม ในการส่งเสริมความแม่นยำในการทำนายประเภทของที่อยู่เว็บไซต์ให้แก่แบบจำลองการเรียนรู้ของเครื่อง ระหว่างกลุ่มคุณลักษณะทางภาษา และ กลุ่มคุณลักษณะทางเนื้อหานั้น ผลการทำนายของแบบจำลองการเรียนรู้ (Confusion matrix) จะถูกนำมาคำนวนโดยการแทนค่าลงไปในสมการเพื่อหาค่า Accuracy, Precision, recall และ F1 Score และนำผลลัพธ์ที่ได้จากการทดลองในแต่ละกลุ่มมาทำการสรุปและเปรียบเทียบ จากตารางที่ 3.6 เป็นตัวอย่างตาราง Confusion matrix

ตารางที่ 3.6 ตาราง Confusion matrix

	Predicted Negative	**Predicted Positive**
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

จากตาราง Confusion Matrix สามารถสรุปความหมายของค่าต่าง ๆ ได้ดังนี้

- True Positive (TP): จำนวนข้อมูลที่ถูกทำนายว่าเป็น Positive และตรงกับความจริงว่าเป็นPositive
- False Positive (FP): จำนวนข้อมูลที่ถูกทำนายว่าเป็น Positive แต่ตรงกับความจริงว่าเป็น Negative
- False Negative (FN) - จำนวนข้อมูลที่ถูกทำนายว่าเป็น Negative แต่ตรงกับความจริงว่าเป็น Positive
- True Negative (TN) - จำนวนข้อมูลที่ถูกทำนายว่าเป็น Negative และตรงกับความจริงว่าเป็น Negative

โดยในส่วนของการคำนวณหาค่าความถูกต้อง (Accuracy) นั้นคือการคำนวนเพื่อหาสัดส่วนของข้อมูลที่ถูกทำนายได้ถูกต้องทั้งหมด สามารถคำนวณได้โดยแทนค่าลงในสมการ (1)

- $\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$ (1)

สำหรับการคำนวณหาค่าความแม่น (Precision) คือ การคำนวนเพื่อหาสัดส่วนของข้อมูลที่ถูกทำนายว่าเป็น Positive และตรงกับความจริงต่อจำนวนทั้งหมดของข้อมูลที่แบบจำลองทำนายว่าเป็น Positive ซึ่งสามารถคำนวณได้โดยการแทนค่าลงในสมการ (2)

- $\text{Precision} = \frac{TP}{TP+FP}$ (2)

ในส่วนของการคำนวณเพื่อหาค่า Recall หรือ สัดส่วนของข้อมูลที่แบบจำลองทำนายว่าเป็น Positive และ ตรงกับความจริง ต่อจำนวนของข้อมูลจริงที่เป็น Positive สามารถคำนวนโดยใช้สมการ (3)

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

สุดท้ายสำหรับการคำนวณหาค่า F1 Score ซึ่งเป็นค่าที่ใช้ในการประเมินประสิทธิภาพของแบบจำลองการเรียนรู้ในการทำนาย โดยการนำค่า Precision และ Recall มาคำนวณเพื่อหาค่าเฉลี่ยฮาร์โมนิก (Harmonic mean) ที่แสดงถึงความสมดุลระหว่าง Precision และ Recall ของแบบจำลองการเรียนรู้ โดยสามารถคำนวณโดยการแทนค่าลงในสมการ (4)

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{4}$$

บทที่ 4
ผลการวิจัย

การศึกษาและวิจัยในครั้งนี้ เป็นการศึกษาเกี่ยวกับการจำแนกที่อยู่เว็บไซต์เพื่อป้องกันการโจมตีแบบฟิชชิงผ่านทางที่อยู่เว็บไซต์ โดยใช้เทคนิคการเรียนรู้ของเครื่อง โดยจุดประสงค์หลักของการทำการทดลองในครั้งนี้ คือ เพื่อศึกษาและเปรียบเทียบประสิทธิภาพในการส่งเสริมประสิทธิภาพการเรียนรู้ และความแม่นยำในการทำนายที่อยู่เว็บไซต์อันตราย (Malicious URL) ให้แก่ของแบบจำลองการเรียนรู้ของเครื่องระหว่างกลุ่มคุณลักษณะทางภาษา และ กลุ่มคุณลักษณะทางเนื้อหา ซึ่งเป็นกลุ่มคุณลักษณะที่มักปรากฏในงานวิจัยในงานตรวจสอบที่อยู่เว็บไซต์ โดยคุณลักษณะทั้ง 2 กลุ่มนี้ มีความเกี่ยวข้องกับการวิเคราะห์โครงสร้างทางภาษา แต่มีความแตกต่างกันคือ กลุ่มคุณลักษณะทางภาษา (Lexical feature) เป็นการวิเคราะห์โครงสร้างและองค์ประกอบที่ปรากฏบนชื่อที่อยู่เว็บไซต์ ในขณะที่กลุ่มคุณลักษณะทางเนื้อหา (Content-based feature) นั้น เป็นการวิเคราะห์โครงสร้างและองค์ประกอบของหน้าเว็บไซต์ที่เชื่อมต่อกับที่อยู่เว็บไซต์นั้น ๆ ซึ่งหมายถึงสิ่งที่ปรากฏบน JavaScript หรือ โค้ด HTML

ในการทดลองครั้งนี้ เทคนิค Recursive Feature Elimination (RFE) ถูกนำมาใช้ในการคัดเลือกคุณลักษณะที่สำคัญ 10 คุณลักษณะ จากกลุ่มคุณลักษณะทางภาษา และ จากกลุ่มคุณลักษณะทางเนื้อหา โดยคุณลักษณะสำคัญจากแต่ละกลุ่ม จะถูกนำมาใช้ในการสร้างแบบจำลองการเรียนรู้ โดยใช้ อัลกอริทึมป่าไม้แบบสุ่ม (Random Forest) เพื่อเปรียบเทียบประสิทธิภาพระหว่างกลุ่มคุณลักษณะทั้ง 2 กลุ่ม จึงได้ทำการสร้างโมเดลขึ้นมาจำนวน 2 โมเดล โดยโมเดลที่ 1 ใช้ 10 คุณลักษณะสำคัญ 10 อันดับแรกของกลุ่มคุณลักษณะทางภาษา และ แบบโมเดลที่ 2 ใช้ 10 คุณลักษณะสำคัญ 10 อันดับแรก จากกลุ่มคุณลักษณะทางเนื้อหา หลังจากนั้นนำผลลัพธ์จากการทำนาย (Confusion Matrix) ของการทดลองแต่ละโมเดล มาคำนวนหาค่า Accuracy (1), Precision (2), Recall (3) และ F1 Score (4) โดยการแทนค่าลงไปในสมการ (1) (2) (3) (4) และนำผลลัพธ์ที่ได้มาเปรียบเทียบสรุปผลการทดลอง โดยตารางที่ 4.1 เป็นตารางผลการทำนายแบบ Confusion matrix ของทั้งสองกลุ่ม

- $\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$ (1)
- $\text{Precision} = \frac{TP}{TP+FP}$ (2)
- $\text{Recall} = \frac{TP}{TP+FN}$ (3)
- $\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$ (4)

ตารางที่ 4.1 ผลการทำนายที่อยู่เว็บไซต์ของแต่ละกลุ่มคุณลักษณะ

	Predicted Negative	**Predicted Positive**
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)
โมเดลที่ 1	**Predicted Benign**	**Predicted Malicious**
	5,650	2,412
	2,914	5,148
โมเดลที่ 2	**Predicted Benign**	**Predicted Malicious**
	8,039	23
	317	7,745

สำหรับค่าต่างๆ ที่ปรากฏในตารางที่ 4.1 สามารถสรุปความหมายได้ดังนี้

- True Positive (TP): โมเดลทำนายว่า URL นั้นๆ เป็น URL ที่ไม่ปลอดภัย (Malicious) และ URL นั้น ๆ เป็น URL ที่ไม่ปลอดภัยจริง
- True Negative (TN): โมเดลทำนายว่า URL นั้นๆ เป็น URL ที่ปลอดภัย (Benign) และ URL นั้น ๆ เป็น URL ที่ปลอดภัยจริง
- False Positive (FP): โมเดลทำนายว่า URL นั้นๆ เป็น URL ที่ไม่ปลอดภัย แต่จริง ๆ URL นั้น ๆ เป็น URL ที่ปลอดภัย
- False Negative (FN): โมเดลทำนายว่า URL นั้นๆ เป็น URL ที่ปลอดภัย แต่จริง ๆ URL นั้น ๆ เป็น URL ไม่ปลอดภัย

โดยผลจากการนำผลจาก Confusion matrix มาแทนค่าลงในสมการที่ (1) (2) (3) (4) สามารถสรุปค่าต่างๆ ได้ตามตารางที่ 4.2

ตารางที่ 4.2 ตารางสรุปผลการทดลอง

โมเดล	Accuracy	Precision	Recall	F1 Score
โมเดลที่ 1	0.6696	0.6809	0.6385	0.6590
โมเดลที่ 2	0.9789	0.9970	0.9606	0.9785

จากตารางที่ 4.2 สามารถสรุปได้ว่าการทดลองในครั้งนี้ โมเดลที่ 1 ที่ใช้คุณลักษณะทางภาษาในการเรียนรู้เพื่อจำแนกที่อยู่เว็บไซต์อันตราย (Malicious URL) มีค่าความถูกต้องในการทำนายโดยรวม (Accuracy) อยู่ที่ 67% สำหรับค่า Precision อยู่ที่ 68% และค่า Recall หรืออัตราส่วนของที่อยู่เว็บไซต์ที่เป็นอันตรายทั้งหมดที่แบบจำลองการเรียนรู้สามารถจำแนกได้ถูกต้อง อยู่ที่ 64% และสุดท้าย ค่า F1 Score หรือค่าเฉลี่ยความสมดุลระหว่าง Precision และ Recall อยู่ที่ 66% ในขณะที่โมเดลที่ 2 ที่ใช้ คุณลักษณะสำคัญ 10 อันดับแรกจากกลุ่มคุณลักษณะทางเนื้อหา มีค่าความถูกต้องในการทำนายทั้งหมด หรือ ค่า Accuracy อยู่ที่ 98% ค่า Precision อยู่ที่ 100% ค่า Recall อยู่ที่ 96% และค่าเฉลี่ยความสมดุลระหว่างค่า Precision และ ค่าRecall (F1 Score) อยู่ที่ 98%

ซึ่งจากการทดลองในครั้งนี้สามารถสรุปได้ว่าการใช้คุณลักษณะทางเนื้อหา สามารถช่วยส่งเสริมประสิทธิภาพในการจำแนกและทำนายเว็บไซต์อันตราย (Malicious URL) ให้กับแบบจำลองการเรียนรู้ของเครื่องได้ดีกว่า เมื่อเทียบกับคุณลักษณะทางภาษา โดยโมเดลเรียนรู้ป่าสุ่ม (Random forest) ที่ใช้คุณลักษณะทางเนื้อหานั้นมีความแม่นยำสูงถึง 98% นอกจากนี้ผลการทำนาย Confusion matrix จากตารางที่ 4.1 ก็ยังแสดงให้เห็นว่า โมเดลที่ใช้คุณลักษณะทางเนื้อหานั้น มีประสิทธิภาพในการทำนายที่อยู่เว็บไซต์อันตราย (Malicious URL) ได้ดีกว่าโดยจากจำนวนของตัวอย่างที่อยู่เว็บไซต์อันตราย (Malicious URL) ทั้งหมดจำนวน 8,062 ที่อยู่เว็บไซต์ แบบจำลองได้ทำทายผิดว่าเป็นเป็นที่อยู่เว็บไซต์ปลอดภัย (False Negative) จำนวน 317 ที่อยู่เว็บไซต์ ซึ่งคิดเป็นร้อยละ 3.93 จากจำนวนที่อยู่เว็บไซต์อันตรายทั้งหมด ในขณะที่แบบจำลองการเรียนรู้ที่ใช้คุณลักษณะในกลุ่มคุณลักษณะทางภาษานั้น มีค่า False Negative อยู่ที่ 2,914 ซึ่งคิดเป็นร้อยละ 36.14 จากจำนวนตัวอย่างที่อยู่เว็บไซต์อันตรายทั้งหมด

บทที่ 5
ข้อสรุปและเสนอแนะ

งานวิจัยภายใต้หัวข้อ “การป้องกันการตกเป็นเหยื่อจากการฟิชชิงผ่านทางที่อยู่เว็บไซต์” ฉบับนี้นั้น เป็นการศึกษาเกี่ยวกับการใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ในการจำแนก (Classification) ที่อยู่เว็บไซต์ออกเป็น 2 ประเภท คือ ที่อยู่เว็บไซต์อันตราย (Malicious URL) และ ที่อยู่เว็บไซต์ปลอดภัย (Benign URL) เพื่อป้องกันการโจมตีแบบฟิชชิงผ่านทางที่อยู่เว็บไซต์ โดยในงานวิจัยนี้ ผู้วิจัยกำหนดขอบเขตการศึกษาเกี่ยวกับคุณลักษณะ โดยเลือกศึกษาเฉพาะกลุ่มคุณลักษณะที่เกี่ยวข้องกับการวิเคราะห์โครงสร้าง และองค์ประกอบทางภาษา คือ คุณลักษณะในกลุ่มคุณลักษณะทางภาษา (Lexical feature) และ คุณลักษณะในกลุ่มคุณลักษณะทางเนื้อหา (Content-based feature) ซึ่งมักปรากฏอยู่ในงานวิจัยที่เกี่ยวข้องกับที่ศึกษาเกี่ยวกับการจำแนกที่อยู่เว็บไซต์อันตราย (Malicious URL) ซึ่งกลุ่มคุณลักษณะทางภาษานั้น เป็นการวิเคราะห์จากโครงสร้างทางภาษาที่ปรากฏอยู่ตามที่อยู่เว็บไซต์นั้น ๆ เช่น ความยาวของที่อยู่เว็บไซต์ จำนวนอักขระพิเศษ จำนวนตัวอักษรเล็ก-ใหญ่ ที่ปรากฏบนชื่อที่อยู่เว็บไซต์ ในขณะที่ กลุ่มคุณลักษณะทางเนื้อหาเป็นการวิเคราะห์จากโครงสร้างทางภาษาที่ปรากฏในโค้ด HTML หรือ JavaScript เช่น ความยาวของโค้ด จำนวนของ Hyper link และ จำนวนของ <tag> ที่ปรากฏ เป็นต้น

โดยวัตถุประสงค์ในการศึกษาครั้งนี้ คือ เพื่อทำความเข้าใจเกี่ยวกับลักษณะของคุณลักษณะทางภาษา (Lexical feature) และ คุณลักษณะทางเนื้อหา (Content-based feature) และเพื่อเปรียบเทียบประสิทธิภาพระหว่างกลุ่มคุณลักษณะดังกล่าว ในการส่งเสริมการเรียนรู้ในการจำแนกที่อยู่เว็บไซต์อันตราย (Malicious URL) ให้แก่แบบจำลองการเรียนรู้ของเครื่อง โดยเนื้อหาในบทนี้นั้นจะประกอบไปด้วย 2 ส่วน คือ ส่วนที่เกี่ยวข้องกับการสรุปผลการทดลอง และ ส่วนที่เป็นข้อเสนอแนะ

5.1 การสรุปผลและอภิปรายผลการวิจัย

การทดลองในครั้งนี้ ดำเนินการทดลองโดยใช้ชุดข้อมูล “Malicious and Benign Webpages Dataset” ที่รวบรวมโดย A.K.Singh [31] ซึ่งข้อมูลดังกล่าว มีลักษณะเป็นชุดข้อมูลที่มีความไม่สมดุลกันระหว่างจำนวนของข้อมูล 2 ประเภท คือ ที่อยู่เว็บไซต์อันตราย (Malicious URL) และ ที่อยู่เว็บไซต์ปลอดภัย (Benign URL) เพื่อแก้ปัญหาดังกล่าว ผู้วิจัยจึงใช้ เทคนิค Random Undersampling เพื่อช่วยลดช่องว่างระหว่างจำนวนของข้อมูล 2 ประเภท หลังจากนั้นทำการสกัดคุณลักษณะ 2 กลุ่ม คือ คุณลักษณะทางภาษา และ คุณลักษณะทางเนื้อหา โดยใช้ข้อมูลที่อยู่เว็บไซต์ที่ปรากฏในชุดข้อมูล และ ใช้เทคนิค Recursive Feature Elimination (RFE) ในการช่วยคัดเลือก 10 คุณลักษณะที่สำคัญ จากแต่ละกลุ่ม มาทำการทดลองโดยใช้อัลกอริทึมป่าไม้แบบสุ่ม (Random Forest) โดยทำการสร้างแบบจำลองการเรียนรู้ของเครื่องจำนวน 2 แบบจำลอง โดยแบบจำลองที่ 1 เป็นแบบจำลองการเรียนรู้ของเครื่องที่ ใช้คุณลักษณะสำคัญจำนวน 10 อันดับแรกของกลุ่มคุณลักษณะทางภาษา (Lexical feature) เป็นองค์ประกอบ และแบบจำลองที่ 2 เป็นแบบจำลองการเรียนรู้ที่ใช้คุณลักษณะสำคัญ 10 อันดับแรกจากกลุ่มคุณลักษณะทางเนื้อหา (Content-based feature)

โดยการประเมินผลการทดลองนั้น ผู้วิจัยได้นำผลการทำนาย (Confusion Matrix) ของแต่ละแบบจำลอง มาทำการคำนวณ เพื่อหาค่า Accuracy, Precision, Recall และ F1 Score โดย ผลการคำนวณออกมาว่า แบบจำลองการเรียนรู้ป่าไม้แบบสุ่มที่ใช้คุณลักษณะสำคัญ 10 อันดับแรกจากกลุ่มคุณลักษณะทางเนื้อหา (Content-based feature)นั้น ส่งผลให้แบบจำลองป่าไม้แบบสุ่ม มีความแม่นยำในการทำนายที่อยู่เว็บไซต์อันตรายสูงกว่า แบบจำลองการเรียนรู้ที่ใช้คุณลักษณะทางภาษา (Lexical feature) โดยค่า Accuracy ของ แบบจำลองการเรียนรู้ที่ใช้คุณลักษณะทางเนื้อหา (Content-based feature) นั้น สูงถึง 98% ในขณะที่ คุณลักษณะทางภาษา (Lexical feature) นั้นได้เพียง 67% เท่านั้น ทั้งนี้หากวิเคราะห์จากผลการทำนาย (Confusion Matrix) จะพบว่าการทดลองโดยใช้คุณลักษณะทางเนื้อหามีค่า False Negative (จำนวนตัวอย่างที่อยู่เว็บไซต์อันตราย ที่แบบจำลองทายผิดว่าเป็นเว็บไซต์ปลอดภัย) อยู่เพียง 317 ที่อยู่เว็บไซต์ จากจำนวนตัวอย่างที่อยู่เว็บไซต์อันตรายทั้งหมด 8,062 ที่อยู่เว็บไซต์ โดยคิดเป็นร้อยละ 3.93 จากจำนวนตัวอย่างเว็บไซต์อันตรายทั้งหมด เมื่อเทียบกับแบบจำลองการเรียนรู้ที่ใช้คุณลักษณะทางภาษานั้น มีค่า False Negative อยู่ที่ 2,932 คิดเป็นร้อยละ 36.36 จากตัวอย่างเว็บไซต์อันตรายทั้งหมด

จากผลการศึกษาและทดลองในครั้งนี้พบว่า แบบจำลองการเรียนรู้ที่ใช้คุณลักษณะในกลุ่มคุณลักษณะทางเนื้อหา (Content-based feature) มีประสิทธิภาพในการเรียนรู้ และ ความสามารถในการจำแนกที่อยู่เว็บไซต์อันตรายได้ดีกว่า แบบจำลองการเรียนรู้ที่ใช้กลุ่มคุณลักษณะทางภาษา (Lexical feature)

นอกจากนี้เมื่อเปรียบเทียบกับงานวิจัยที่ศึกษาโดย A. Saleem Raja และ คณะที่ทำการสร้างแบบจำลอง Light weighted สำหรับจำแนกที่อยู่เว็บไซต์อันตราย โดยใช้ คุณลักษณะทางภาษา (Lexical feature) จำนวน 20 คุณลักษณะ อย่างไรก็ตามถ้าเปรียบเทียบโดยมองเรื่อง Light weighted แบบจำลองการเรียนรู้ที่ใช้ เพียง 10 คุณลักษณะจากกลุ่มคุณลักษณะทางเนื้อหา (Content-based feature) จากงานวิจัยนี้นั้น น่าจะตอบโจทย์ในส่วนของการสร้างแบบจำลองการเรียนรู้ Light weighted มากกว่า

5.2 ข้อเสนอแนะ

สำหรับงานวิจัยในครั้งนี้ ด้วยข้อจำกัดด้านอุปกรณ์ที่ใช้ในการทดลอง จึงมีความจำเป็นที่จะต้องลดจำนวนข้อมูลลงก่อนที่จะทำการสกัดคุณลักษณะ โดยวิธี Random undersampling เป็นวิธีการที่ช่วยลดจำนวนข้อมูลในกลุ่มที่มีมากกว่า ให้ลดลงเท่ากับกลุ่มที่มีจำนวนน้อยกว่า ซึ่งการลดลงแบบสุ่มนั้น อาจทำให้ข้อมูลสำคัญบางส่วนถูกลบไป ซึ่งหากทำการสกัดคุณลักษณะก่อนที่จะทำการลดจำนวนข้อมูลอาจจะช่วยให้ลดโอกาสในการสูญเสียข้อมูลที่สำคัญไป

นอกจากนี้หากทำการศึกษาโดยสกัดคุณลักษณะจากชุดข้อมูลที่แตกต่างกัน ก็จะสามารถช่วยให้สามารถเข้าใจลักษณะของข้อมูลต่าง ๆ ได้ดีขึ้น สามารถปรับปรุงแบบจำลองให้มีประสิทธิภาพสูงขึ้น และสามารถใช้งานในสถานการณ์จริงได้ดีขึ้น

การเลือกใช้คุณลักษณะเพียงกลุ่มใดกลุ่มหนึ่งอาจจะช่วยลดความซับซ้อน และ ลดการใช้ทรัพยากร (Light weighted method) ได้ แต่อย่างไรก็ตามเพื่อจะพัฒนาระบบที่สามารถป้องกันการโจมตีแบบฟิชชิงผ่านทางที่อยู่เว็บไซต์ได้มีประสิทธิภาพที่สูงขึ้น จะต้องพิจารณาถึงองค์ประกอบอื่น ๆ อีก เช่น ข้อจำกัดด้านทรัพยากร รวมถึงรูปแบบการโจมตีที่กำลังเผชิญ ซึ่งการเลือกใช้คุณลักษณะให้เหมาะกับปัญหาจะช่วยให้สามารถสร้างแบบจำลองการเรียนรู้ที่มีประสิทธิภาพในการจัดการกับปัญหาได้ดียิ่งขึ้น

บรรณานุกรม

บรรณานุกรม

[1] การเงินและธนาคาร, *“แคสเปอร์สกี้ เผยภัยฟิชชิงไทยติดอันดับ 3 อาเซียน,”* [Online]. Available: https://moneyandbanking.co.th/2023/51280/ [Accessed: 30 สิงหาคม 2566]

[2] Thaipost,*“พบคนไทยตกเป็นเหยื่อภัยไซเบอร์ เสียหายกว่า 40,000 ล้านบาท,”* thaipost.net [Online]. Available: https://www.thaipost.net/criminality-news/397937/. [Accessed: 30 สิงหาคม 2566].

[3] M. Aljabri et al., “Detecting malicious URLs using machine learning techniques: Review and research directions,” *IEEE Access*, vol. 10, pp. 121395-121417, November 2022.

[4] S. Asiri et al., “A survey of Intelligent detection designs of HTML URL phishing attacks,” in *IEEE Access*, vol. 11, pp. 6421-6443, January 2023.

[5] M. Aljabri and S.Mirza, “Phishing attacks detection using machine learning and deep learning models,” *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)* Riyadh, Saudi Arabia, March 1-3, 2022, pp.175-180.

[6] Y. Wang, “Malicious URL detection an evaluation of feature extraction and machine learning algorithm,” *Highlights in Science, Engineering and Technology*, vol.23, pp.117-123, December 2022.

[7] Z. Alkhalil et al., “Phishing attacks: A recent comprehensive study and a new anatomy,” *Frontiers of Computer Science*, vol.3, pp.563060, March 2021.

[8] N. Puri et al., “Application of ensemble machine learning models for phishing detection on web networks,” *Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, Sonepat, India, July 8-9, 2022, pp. 296-303.

[9] S.Mishra and D.Soni, “Smishing detector: A security model to detect smishing through SMS content analysis and URL behavior analysis,” *Future Generation Computer Systems-the international Journal of eScience*, vol. 108, pp. 803-815, July 2020.

[10] A. Zamir et al., "Phishing web site detection using diverse machine learning algorithms," *Electronics Library*, vol. 38, no. 1, pp. 65-80, January 2020.

[11] A. S. Raja et al., "Structural analysis of URL for malicious URL detection using machine learning," *JOAASR*, vol.5, pp28-41, July 2023.

[12] M. Binjubeir, et al., "Comprehensive survey on big data privacy protection," *IEEE Access*, vol. 8, pp.200067-20079, December 2019.

[13] A. Basit et al., "A comprehensive survey of AI-enabled phishing attacks detection techniques," *Telecommunication Systems*, vol. 76, pp. 139-154, October 2020.

[14] S. Ozgur et al., "Machine learning based phishing detection from URLs," *Expert Systems with Applications*, vol. 117, pp.345-357, January 2019.

[15] A. Krishna et al., "Phishing detection using machine learning based URL analysis: A survey," *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, pp. 156-161, August 2021.

[16] R. Zieni et. al., "Phishing or not phishing? A survey on the detection of phishing websites," *IEEE Access*, vol. 11, pp. 18499-18519, February 2023.

[17] L. Tang and Q. H. Mahmoud, "A survey of machine learning-based solutions for phishing website detection," *Machine Learning and Knowledge Extraction*, vol. 3, pp.672-694, August 2021.

[18] M. Alshira and M. A. Fawa'reh., "Detecting phishing URLs using machine learning & lexical feature-based analysis," *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)*, vol.9, no.4, pp.5828-5837, August 2020.

[19] G. Robots, "*Extracting Feature Vectors From URL Strings For Malicious URL Detection,"* [Online]. Available: https://towardsdatascience.com/extractingfeature-vectors-from-url-strings-for-malicious-url-detection-cbafc24737a. [Accessed: November 11, 2023].

[20] P. Yang et al., "Phishing website detection based on multidimensional features driven by deep learning," *IEEE Access*, vol. 7, pp. 15196-15209, January 2019.

[21] M. Elsadig et al., "Intelligent deep machine learning cyber phishing URL detection based on bert features extraction," *in Electronics*, vol.11, no.22, pp.3647 November 2022.

[22] A. S. Raja et al., “Lexical features based malicious URL detection using machine learning techniques,” *Materials Today: Proceedings*, vol. 47, pp. 163-166, April 2021.

[23] W.Faiq and N.G.M. Jameel, “Malicious URL detection tree-based lexical features selection and multilayer perceptron model,” *in UHD Journal of Science and Technology*, vol.6, pp.105-116, November 2022.

[24] B.Altay et al., “Context-sensitive and keyword density based supervised machine learning techmiques for malicious webpage detection,” *in Soft Computing*, vol.23, pp.4177-4191, February 2018.

[25] S.Kumi et al., “Malicious URL detection based on associative classification,” *in Entropy*, vol. 23, no.2, pp.182, January 2021.

[26] J. Liu et al., “Detecting web spam based on novel features from web page source code,” *in Hindawi*, vol. 2020, pp.6662166, December 2020.

[27] P.V. Rao et al., “Detection of malicious uniform resource locator,” *International Journal of Recent Technology and Engineering*, vol.8, pp.41-47, July 2019.

[28] M.N. Alam et al., “Phishing attacks detection using machine learning approach,” *2020 Third International Conference on Smart System and Inventive Technology (ICSSIT)*, Tirunelveli, India, August 20-22, 2020, pp. 1173-1179.

[29] M. Aljabri et al., “An assessment of lexical, network, and content-based features for detecting malicious URLs using machine learning and deep learning models,” *in Hindawi,* vol.2022, pp.3241216, August 2022.

[30] C. D. Xuan et al., “Malicious URL detection based on machine learning,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 1, pp.148-153, January 2020.

[31] A.K.Singh, “Malicious and benign webpages dataset,” in *Data in Brief*, vol.32, pp.106304, October 2020.

[32] T. Oransirikul and H. Takada, “Investigating feature selection for predicting the number of bus passengers by monitoring wi-fi signal activity from mobile devices,” *in Advances in Intelligent Systems and Computing*, vol. 1151, pp. 333-344, March 2020.

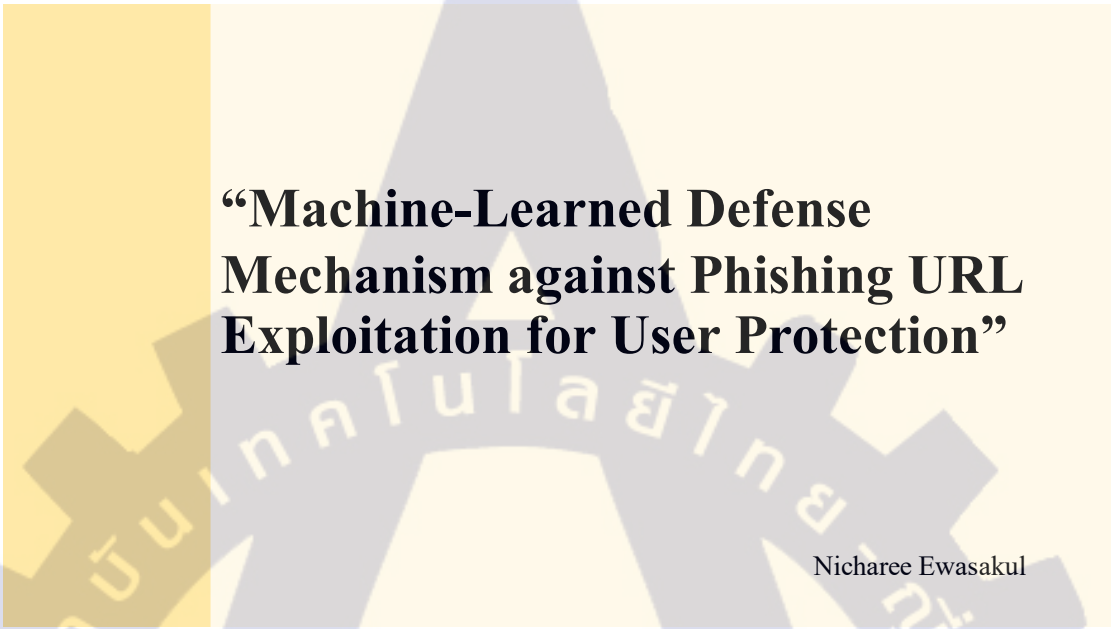
ภาคผนวก

ภาคผนวก ก.

เอกสารประกอบการนำเสนอการสอบป้องกันวิทยานิพนธ์

เอกสารประกอบการนำเสนอการสอบป้องกันวิทยานิพนธ์

1. หน้าปกเอกสารประกอบการนำเสนอ



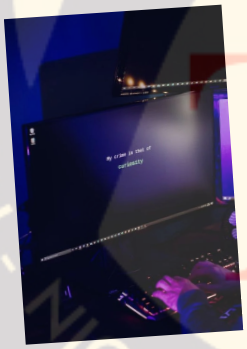
2. หน้าชี้แจงองค์ประกอบของการนำเสนอ

3. หน้าปกของส่วน Introduction

Introduction

4. หน้าเอกสารประกอบการอธิบายเกี่ยวกับการฟิชชิง

Phishing Attack



- A type of cyber attack where attackers deceive individuals into disclosing sensitive information.
- Using social engineering tactics to manipulate targets.
- Can result in financial losses, data breaches, reputational damage, and operational disruption

4

5. หน้าเอกสารประกอบการยกตัวอย่างรูปแบบการฟิชชิง

Types of Phishing

There are several types of Phishing such as

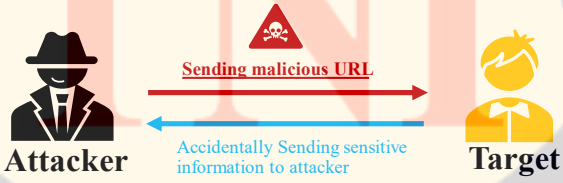
- Email Phishing
- Smishing
- Vishing
- Spear Phishing
- Whaling
- Clone Phishing

5

6. หน้าเอกสารอธิบายเกี่ยวกับการฟิชชิงผ่านทางที่อยู่เว็บไซต์

URL Phishing

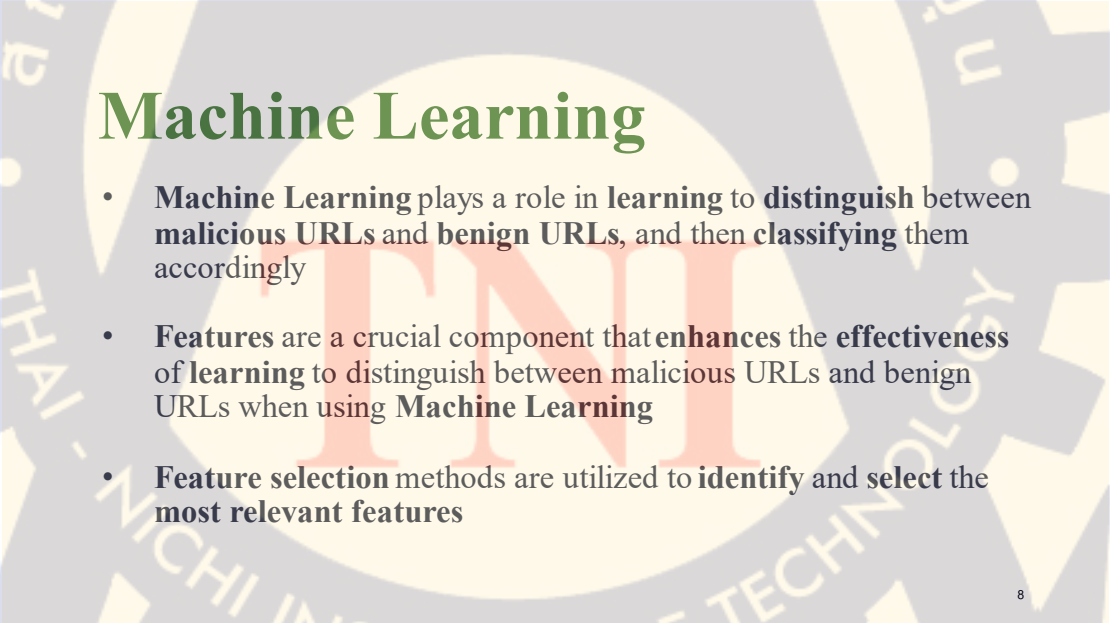
- **a type of cyber attack where attackers create a fake website that mimics a legitimate one. The goal is to deceive users into entering their personal information, such as login credentials or financial data**
- **URLs** are used as **bait** to **phish** individuals



6

7. หน้าเอกสารประกอบการอธิบายเกี่ยวกับวิธีการในการป้องกันการโจมตีแบบฟิชชิงผ่านทางที่อยู่เว็บไซต์

8. หน้าเอกสารประกอบการอธิบายเกี่ยวกับเทคนิคการเรียนรู้ของเครื่อง



9. หน้าเอกสารประกอบการอธิบายเกี่ยวกับคำศัพท์

Malicious URLs VS Benign URLs

- **Malicious URLs** : refer to web addresses designed for harmful activities such as phishing attacks or the spread of malware
- **Benign URLs** : refer to web addresses that are safe and intended for legitimate purposes

9

10. หน้าเอกสารแสดงวัตถุประสงค์ของงานวิจัย

Research Objective

- To examine the characteristics and behaviors of lexical and content-based features
- To compare the effectiveness of lexical features and content-based features in enhancing machine learning effectiveness for classifying malicious URLs and benign URLs

10

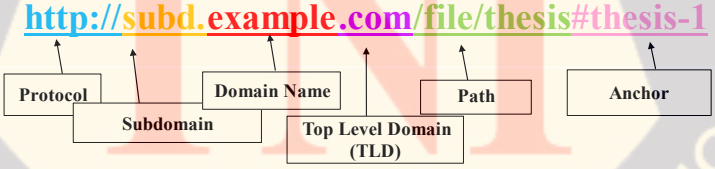
11. หน้าปกส่วนที่ 2 สรุปงานวิจัยที่เกี่ยวข้อง

Related work

12. หน้าเอกสารอธิบายเกี่ยวกับ Lexical feature

Lexical Feature group

Lexical features of URLs refer to the **observable** structural and **surface-level** attributes including **URL length**, **domain information**, and the **presence of special characters**.



12

13. หน้าเอกสารแสดงตัวอย่างของคุณลักษณะ ในกลุ่ม Lexical feature

Lexical Feature Group Example

Feature	Features Category
URL_len	Length
Double-slashes_in_URL	Count
At@_in_URL	Count
Https_in_URL	Count
Numbers_in_URL	Count
Lower_case_letters_in_URL	Count
Upper_case_letters_in_URL	Count
Special_char_in_URL	Count
English_words_in_URL	Count
Avg_English_word_len_in_URL	Length

ที่มา : [1] A. Saleem Raja, R. Vinodini, and A. Kavitha, "Lexical features based malicious URL detection using machine learning techniques, " Materials Today: Proceedings,Volume 47, pp. 163-166, 2021.

13

14. หน้าเอกสารอธิบายเกี่ยวกับ Content-based feature

Content-based Feature group

Content-based features of URLs involve **analyzing** the **actual content** accessible through the web addresses. This includes examining **webpage text**, **HTML structure**, and **semantic elements**, such as **keywords** and **linguistic patterns**.

14

15. หน้าเอกสารแสดงตัวอย่างคุณลักษณะ ในกลุ่ม Content-based feature

Content-based Feature Group Example

Attribute	Description
js_len	Length of JavaScript
Js_obf_len	Length of the obfuscated JavaScript
count_link	Count appearance of JavaScript link() function in content
count_eval	Count appearance of JavaScript eval() function in content
count_exec	Count appearance of JavaScript exec() function in content
count_unescappe	Count appearance of JavaScript unescaped() function in content
count_search	Count appearance of JavaScript search() function in content
count_find	Count appearance of JavaScript find() function in content
Presence_iframe	The presence of the iFrame tag is checked in content
lines_count	The number of lines of the content

ที่มา : [2] M. Aljabri et al., "An Assessment of Lexical, Network, and Content -Based Features for Detecting Malicious URLs Using Machine Learning and Deep Learning Models," in Hindawi, vol.2022, 2022.

15

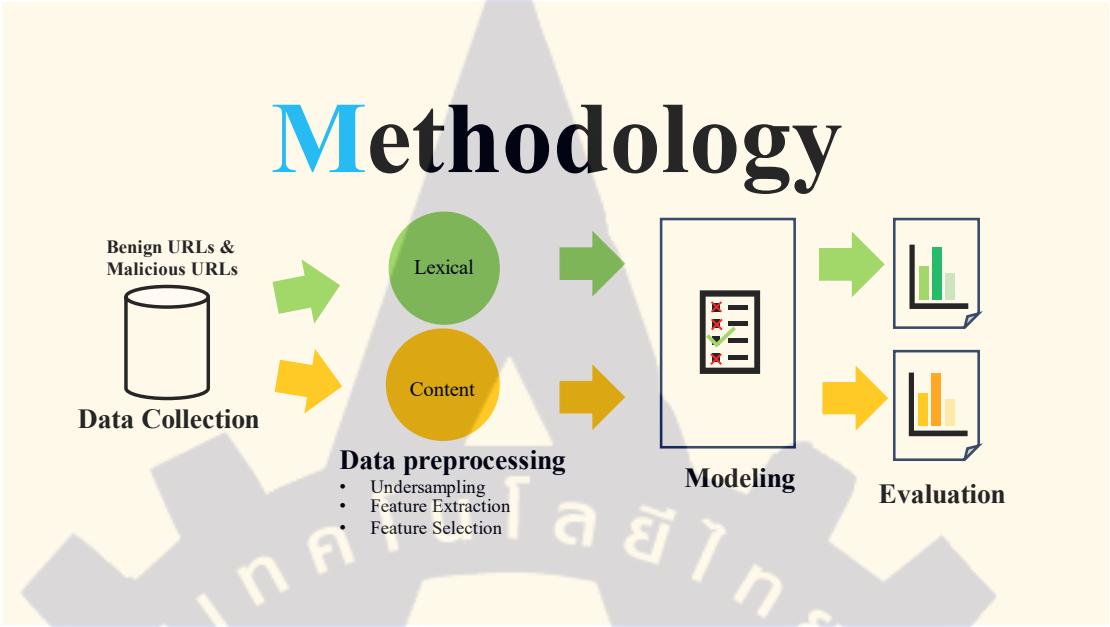
16. หน้าเอกสารประกอบการอธิบายเกี่ยวกับวิธีการเรียนรู้ของเครื่อง

Machine Learning

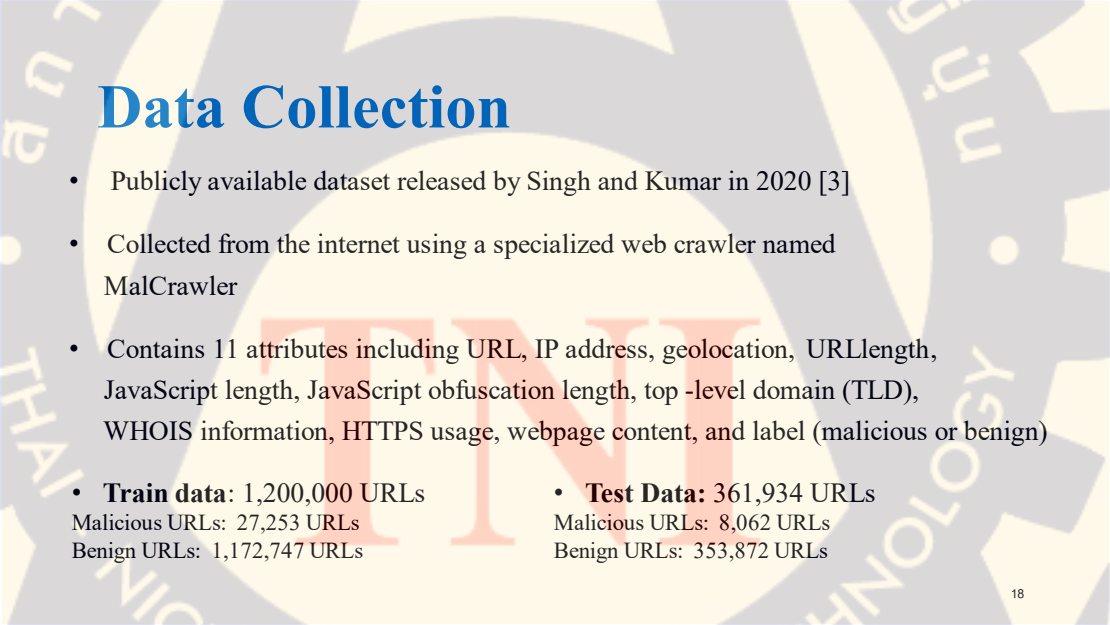
- Traditional machine learning algorithms such as **Support Vector Machine**, **Decision Tree**, **K-Nearest Neighbors**, **Random Forest**
  - B.Altay et al. [5] conducted research using only content features with several models. The highest accuracy achieved was 98% with SVM
  - A. Saleem Raja et al.[6] conducted research using only lexical features with several models. The highest accuracy achieved was 99% with Random forest
- Z. Rasha et al.[4] conducted a survey of phishing detection papers, observing that many papers use multiple algorithms to identify the best performing one, with results showing that **Random Forest tends to outperform other algorithms**

16

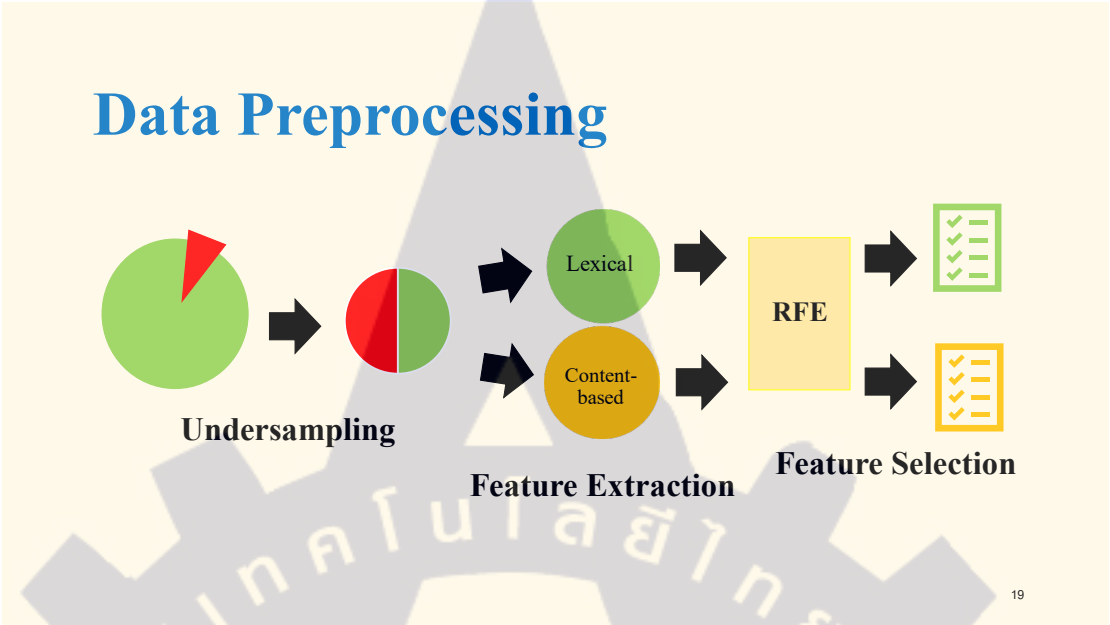
17. หน้าปกส่วนที่ 3 วิธีการดำเนินงาน



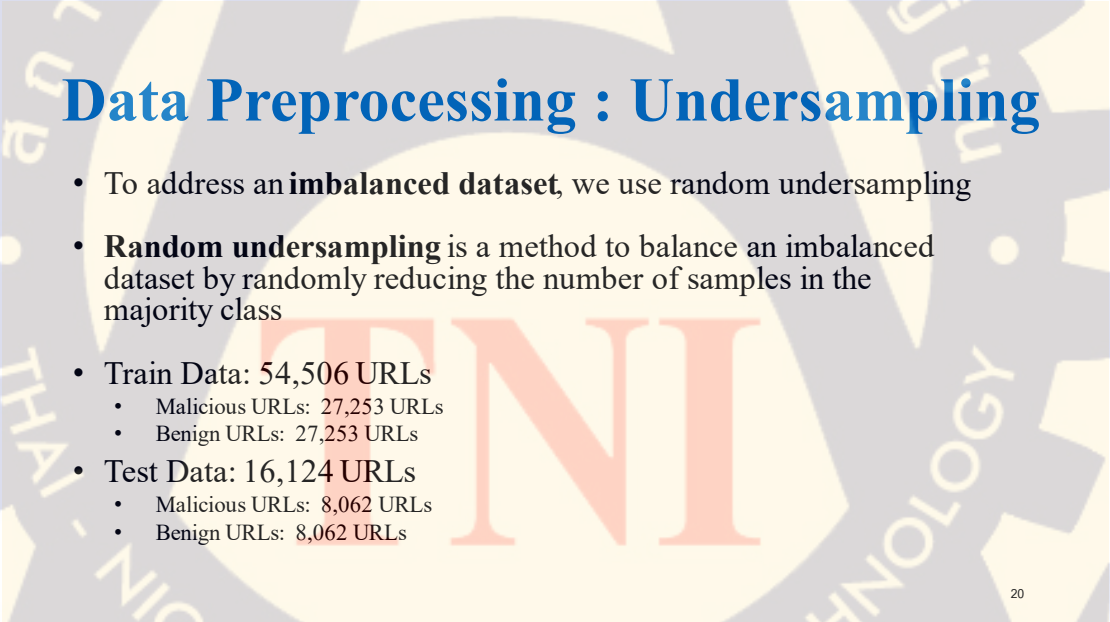
18. หน้าเอกสารอธิบายเกี่ยวกับขั้นตอน Data collectio



19. หน้าเอกสารอธิบายเกี่ยวกับขั้นตอนในส่วนของ Data preprocessing



20. หน้าเอกสารประกอบการอธิบายเกี่ยวกับขั้นตอน Undersampling



21. หน้าเอกสารประกอบการอธิบายขั้นตอน Feature Extraction

Data Preprocessing: Feature Extraction Lexical features

- Lexical features are **characteristics** that **reflect** the **structure** of a **URL**
- The features **include** URL **length**, **presence** of **specific characters** (such as slashes, dots, and symbols), and **patterns observed** in the URL **path** or **host name**
- To extract lexical features, we utilized modules and libraries including
  - '**pandas**' for data manipulation
  - '**math**' for calculations
  - '**re**' for regular expressions
  - '**urllib.parse**' for URL parsing
- In total, this research extracted 19 lexical features

21

22. หน้าเอกสารประกอบการอธิบายขั้นตอน Feature Extraction (ต่อ)

Data Preprocessing: Feature Extraction Content-based features

- Content-based features refer to **characteristics** derived from the **actual content** within a URL, **specifically emphasizing** the **HTML code** and **JavaScript elements**.
- To extract content-based features, we utilized modules and libraries including
  - '**pandas**' for data management
  - '**PyQuery**' for HTML parsing
  - '**string**' for text manipulation
  - '**json**' for handling JSON data.
- In total, this research extracted 24 content-based features.

22

23. หน้าเอกสารประกอบการอธิบายขั้นตอน Feature selection

Data Preprocessing: Feature Selection

Recursive Feature Elimination (RFE)

- RFE identifies the top **10 most relevant features** for model performance by selecting from both the Lexical and content-based feature groups
- Works by training the model on the full feature set, ranking features based on importance, and eliminating the least important features until the desired number is reached

23

24. หน้าเอกสารแสดงคุณลักษณะสำคัญ 10 อันดับแรก ของคุณลักษณะในกลุ่ม Lexical feature

10 selected Lexical Features

Features	Description
1. url_length	The length of URL
2. url_path_length	The length of the path part of a URL
3. url_host_length	The length of the domain name (host) in a URL
4. url_slash_count	The number of slashes ('/')
5. url_dot_count	The number of dots ('.')
6. url_hyphen_count	The number of hyphens ('-')
7. url_question_mark_count	The number of question marks ('?')
8. number_of_digit	The total count of digits (0-9) in the entire URL
9. number_of_parameter	The number of parameters in the URL's query string
10. number_of_subdirectories	The number of subdirectories (folders) in the path part of the URL

24

25. หน้าเอกสารแสดงคุณลักษณะสำคัญ 10 อันดับแรก ของคุณลักษณะในกลุ่ม Content-based feature

10 selected content-based Features

Features	Description
1. js_len	The total length of JavaScript code
2. js_obf_len	The length of obfuscated JavaScript code within the webpage
3. script_to_body_ratio	The ratio of the total length of JavaScript code
4. length_of_html	The total length of the HTML
5. number_of_punctuations	The total count of punctuation marks
6. number_of_distinct_tokens	The number of unique words or tokens
7. average_number_of_tokens_in_sentence	The average number of words (tokens) per sentence
8. number_of_html_tags	The total count of HTML tags used
9. number_of_hyperlinks	The total count of clickable hyperlinks
10. number_of_included_elements	The total count of external resources

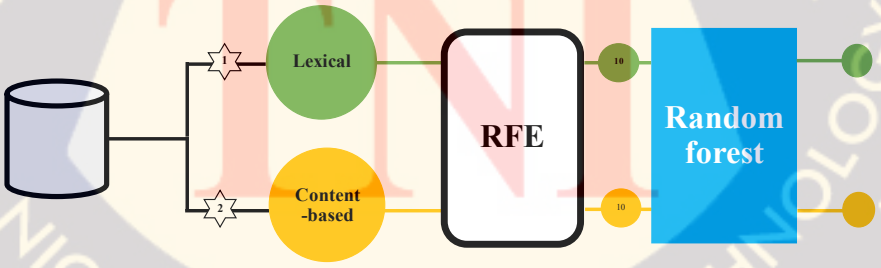
25

26. หน้าเอกสารประกอบการอธิบายการสร้างโมเดล

Modeling

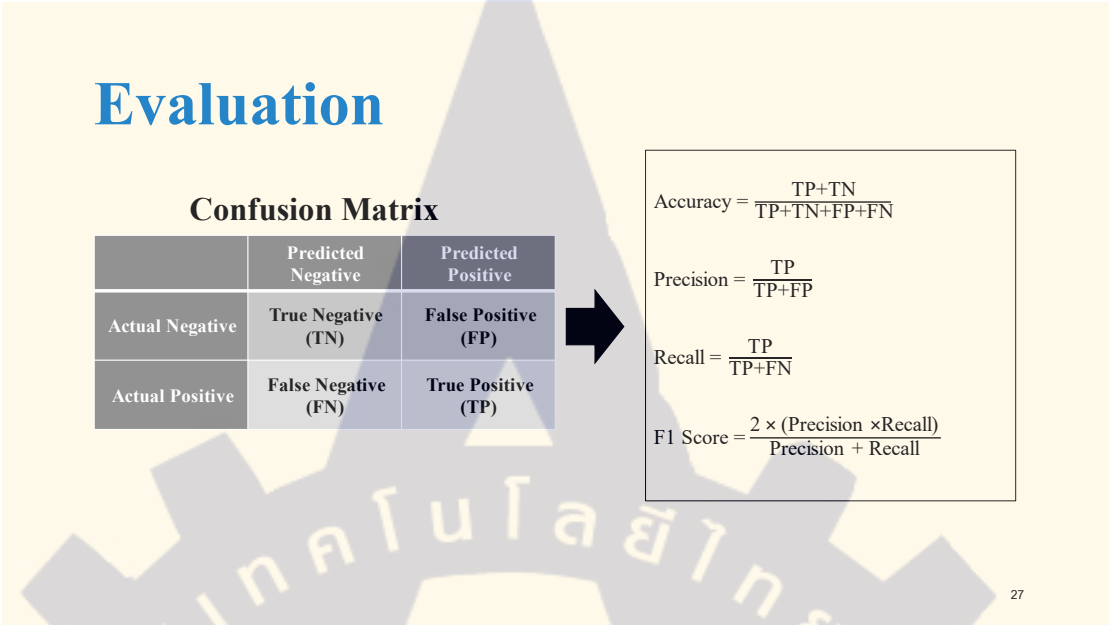
We conducted two experiments divided into twomodels:

- First, utilizing Lexical features with Random Forest
- Second, utilizing content-based features also with Random Forest



26

27. หน้าเอกสารประกอบการอธิบายเกี่ยวกับขั้นตอนการประเมินผล



28. หน้าปกส่วนที่ 4 ผลการทดลอง

29. หน้าเอกสารแสดงผลการทำนาย (Confusion Matrix) ของแต่ละโมเดล

Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

Lexical	Predicted Benign	Predicted Malicious
	5,650	2,412
	2,914	5,148
Content-based	**Predicted Benign**	**Predicted Malicious**
	8039	23
	317	7745

29

30. หน้าเอกสารแสดงผลการทดลองของโมเดลที่ 1 ที่ใช้ Lexical feature

Result
Random forest with Lexical feature

	Accuracy	Precision	Recall	F1 Score
Lexical	0.67	0.68	0.64	0.66

- Achieved accuracy rate of 67%.
- While the precision rate of 68%, the recall rate of 64%.
- F1 Score of 66% reflects the balanced performance considering both precision and recall.

30

31. หน้าเอกสารแสดงผลการทดลองของโมเดลที่ 2 ที่ใช้ Content-based feature

Result

Random forest with content-based feature

	Accuracy	Precision	Recall	F1 Score
Content-based	0.98	0.99	0.96	0.98

- Achieved high accuracy of 98%
- The precision of 99% and recall of 96%, indicating excellent performance in identifying malicious URLs based on content features.
- F1 Score of 98% highlights the effectiveness of content -based features in the detection process.

31

32. หน้าปกส่วนที่ 5 สรุปผล

Conclusion

33. หน้าเอกสารประกอบการอธิบายเกี่ยวกับการสรุปผลการทดลอง

Conclusion

Lexical Features Performance

- Achieved accuracy rate of 67%
- Correctly identified many benign URLs
- Missed a significant portion of malicious URLs

Content-based feature Performance

- Achieved highest accuracy rate of 98%, precision, recall, and F1 score
- Identified both benign and malicious URLs accurately
- Achieved **high accuracy** with **minimal false positives**

The experiment shows that using different sets of features leads to different outcomes in how well the model performs

33

34. หน้าเอกสารแสดงเอกสารอ้างอิงที่ปรากฏบนเอกสารการนำเสนอ

References

[1] A. Saleem Raja, R. Vinodini, and A. Kavitha, “Lexical features based malicious URL detection using machine learning techniques,” Materials Today: Proceedings,Volume 47, pp. 163-166, 2021.

[2] M. Aljabri et al., “An Assessment of Lexical, Network, and Content -Based Features for Detecting Malicious URLs Using Machine Learning and Deep Learning Models,” in Hindawi, vol.2022, 2022.

[3] A.K.Singh, “Malicious and Benign Webpages Dataset,” in Data in brief, vol.32, 2020

[4] R. Zieni, L. Massari and M. C. Calzarossa, "Phishing or Not Phishing? A Survey on the Detectionof Phishing Websites," in IEEE Access, vol. 11, pp. 18499 -18519, 2023.

[5] B.Altay, T.Dokeroglu, and A.Cosar, “Context-sensitive and keyword density based supervised machine learning techmiques for malicious webpage detection,” in Soft computing, vol.23, June 2019.

[6] A. Saleem Raja, R. Vinodini, and A. Kavitha, “Lexical features based malicious URL detection using machine learning techniques,” Materials Today: Proceedings, Volume 47, pp. 163 -166, 2021.

35