第一章　緒論

在網際網路如此發達的年代，網路攻擊儼然成為我們最需要防範的事項之一，如果不慎遭受攻擊，可能導致個人的如信用卡號碼、帳號密碼等機密資訊外流，其中一種攻擊手段即為社交工程(social engineering)，常見的手法偽裝成政府機構或是宅配業者等，在郵件或簡訊中放入惡意網址引誘他人點選，而這些惡意網址便會透過各種形式來欺騙使用者，以下列舉幾種常見的詐騙形式:

圖: 偽裝成銀行的簡訊詐騙 (來源:whoscall)

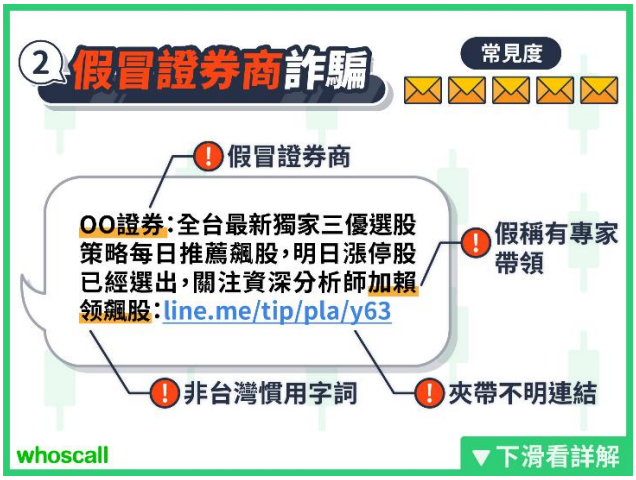

圖: 投資簡訊詐騙 (來源:whoscall)

圖：偽裝成政府的簡訊詐騙 (來源:whoscall)

圖：我實際收到的詐騙簡訊

有鑑於近期實在收到太多這種包含惡意網址的詐騙簡訊，本次的期末報告期望可以透過人工智慧的技術來辨別惡意網址，一方面可以實踐課程所學的技術，另一方面也可以使我更加了解資料科學於資訊安全領域的應用。

第二章　資料介紹

本文選用 Siddhartha, M. (2021)於 Kaggle 所刊登的資料集，此為作者蒐集不同惡意 URL 相關的資料集整合而成，該資料集並無缺失值，共包含兩個欄位: URL 和 label，其中 label 為標記該網站為良性(benign)、竄改(defacement)、釣魚(phishing)、惡意(malware)，特徵的建立上我參考施淳譯(2020)、Rasymas, T.& Dovydaitis, L.(2020)、Ozcan, A., Catal, C., Donmez, E. et al.(2021)產生了共創造了以下 26 個變數，包含:

變數名稱	**說明**
and	The number of "&" in the URL
backslash	The number of "\" in the URL
comma	The number of "," in the URL
dash	The number of "-" in the URL
digit	The number of digits in the URL
domain_len	Total length of the domain
dot	The number of "." in the URL
exclamation	The number of "!" in the URL
hashtag	The number of "#" in the URL
http	The number of "http" in the URL
https	The number of "https" in the URL
ip	Does an ip address in the URL?
mouse	The number of "@" in the URL
percent	The number of "%" in the URL
plus	The number of "+" in the URL
question	The number of "?" in the URL
semicolon	The number of ";" in the URL
shorten	Does the URL contain a shortening service?
slash	The number of "/" in the URL
subdomain_len	Total length of the subdomain
tilde	The number of "~" in the URL

tld	Does the url contain a known TLD ("com","org","net","edu","gov","int","mil","us","ca","cn","fr","au","de","jp","nl","uk","mx","no","ru","br","se","es","co","tw") ?
underline	The number of "_" in the url
url_len	Total length of the url
word_count	The number of words obtained after parsing the URL by special characters
www	Does the url contain "www"?

表: 變數名稱與說明

在 label 的處理上我將其轉為數值型，其中 benign 的 URL 為 0，而 defacement、phishing、malware 的 URL 我全部視為惡意的 URL 並標記為 1:

標籤	說明
0	Benign or safe URL
1	Defacement, phishing, or malware URL

表: 標籤說明

資料集中共有 66%為良性的 URL、34%為惡意的 URL，下圖為原始資料集 label 分布的情形:

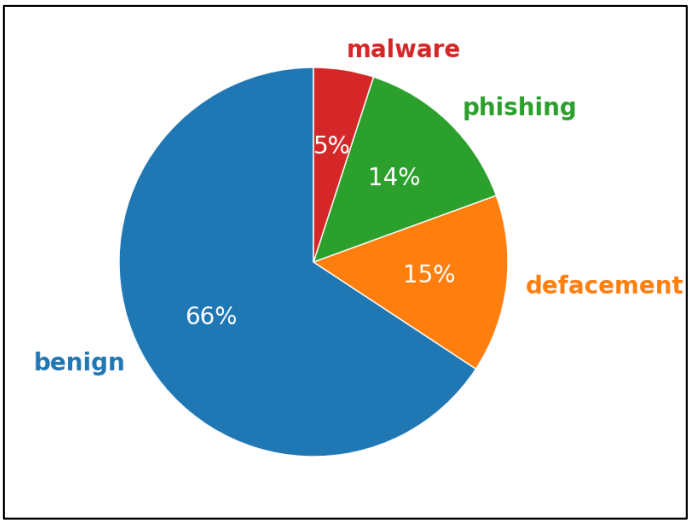

圖: label 分布

第三章　研究方法

本章研究方法共分為四小節，第一節將介紹特徵工程，包含維度縮減以及自然語言處理。第二節敘述本次所採用的機器學習模型。第三節更進一步引進深度學習模型。第四節說明評估模型績效所採用的指標。

第一節　特徵工程

除了第二章所述我參考文獻生成相關的特徵，我希望嘗試使用主成分分析(Principal components analysis，PCA)進行維度下降，探討有沒有做 PCA 對於模型預測結果的影響。此外，我希望能夠將 URL 網址的文字透過自然語言處理中詞嵌入(Word embedding)的方法生成詞向量，並以此作為特徵來觀察應用於模型的表現。

第二節　機器學習模型

由於資料集帶有標籤，適合採用監督式學習演算法，本次選取的模型如下:

- Logistic Regression
- Decision Tree
- Support Vector Machine (SVM)
- Random Forest
- k-Nearest Neighbors (kNN)
- Naive Bayes
- XGBoost

第二節　深度學習模型

在手動生成的特徵和將這些特徵執行 PCA 作為輸入的模型中，選取的深度學習模型如下:

- Deep Neural Network (DNN)

在將 URL 透過自然語言處理轉為詞向量作為輸入的模型中，選取的深度學習模型如下:

- Deep Neural Network (DNN)
- Convolutional Neural Network (CNN)
- Recurrent Neural Networks (RNN)
- Long Short-Term Memory (LSTM)

第三節　績效評估指標

本次選擇的預測績效評估指標如下:

- Accuracy
- Precision
- Recall
- F1-Score

參考文獻

1. 施淳譯（2020）。基於類神經網路之釣魚網站辨識系統，國立中興大學資訊管理學系所碩士論文，台灣台中。
2. Ozcan, A., Catal, C., Donmez, E., & Senturk, B. (2021). A hybrid DNN–LSTM model for detecting phishing URLs. *Neural Computing & Applications*, 1–17.
3. Rasymas, T. & Dovydaitis, L. (2020). Detection of phishing URLs by using deep learning approach and multiple features combinations. *Baltic Journal of Modern Computing*, 8(3), 471–483.
4. Siddhartha, M. (2021) Malicious URLs dataset. Retrieved from https://www.kaggle.com/datasets/sid321axn/malicious-URLs-dataset on May 25, 2022.