認識新軟 AI Anti-Spam

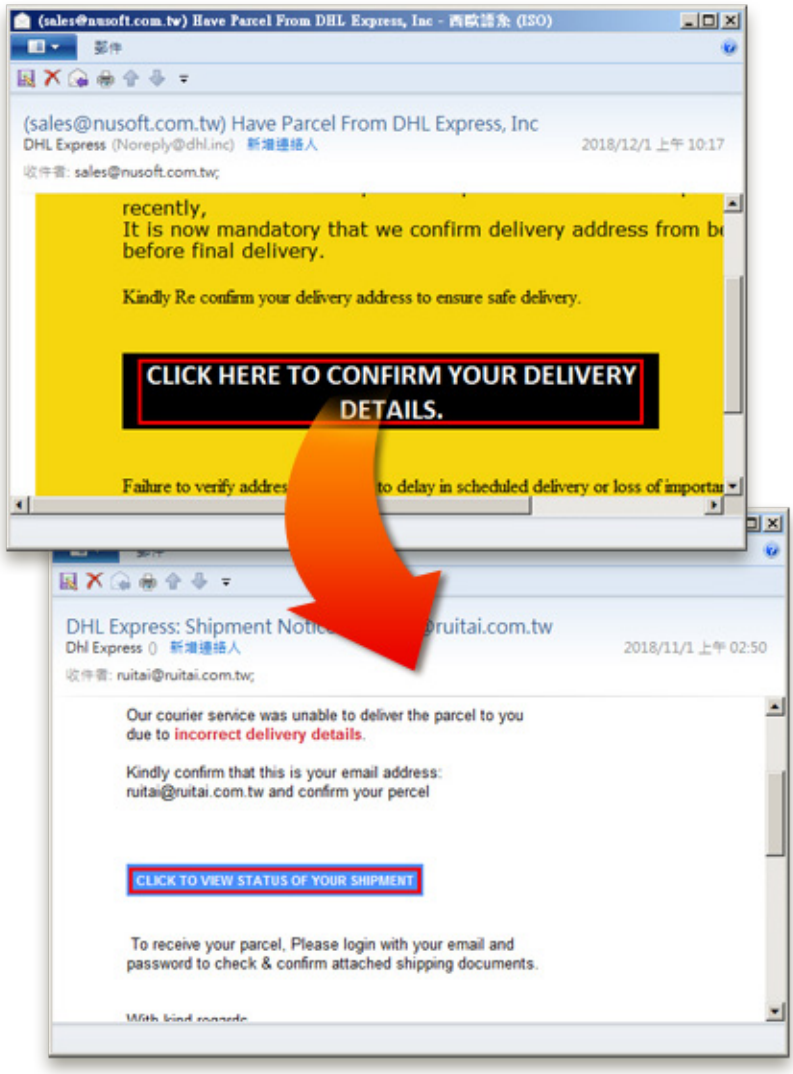
郵件服務是商業活動中訊息往來的重要管道，但長久以來被有心人士做為大量廣告、病毒、…垃圾或惡意資訊的散佈管道，導致正常往來的信件淹沒在一堆公司認為無用甚至有害的郵件中，嚴重影響交易訊息往來要求的時效性。為了因應此狀況，衍生出整合灰名單過濾、個人 / 全體化規則、黑 / 白名單、垃圾郵件特徵、指紋辨識、RBL 資料庫、貝氏過濾、寄件者帳號比對、…多重郵件過濾機制的郵件服務主機，透過許多的規則和特徵碼比對，可以辨識大部分 spam。

有心人士為突破這層層封鎖，寄送的 spam 型態也一直推陳出新，讓郵件服務管理者窮於應付，常常好不容易建立起的新過濾規則，變種 spam 總是能輕易閃過，主要分為下列類型：

(接下頁)

1. 仿造正常信件撰寫方式的釣魚、病毒信件：

釣魚、病毒信件需要被害人開啟檔案或是連結才能達到其目的，所以內容往往與一般商業信件雷同，以欺瞞被害人。這樣的信件因內容用詞常修改的與一般正常信件相同，所以以往過濾機制不易成功判斷。



AI Anti-Spam 會將信件分詞後，與所學 ham 和 spam 字詞資料庫比對。信件內嵌的釣魚網站 URL 字串權重值，高於信件中符合 ham 的字詞權重值，故判斷此類信件為 spam。

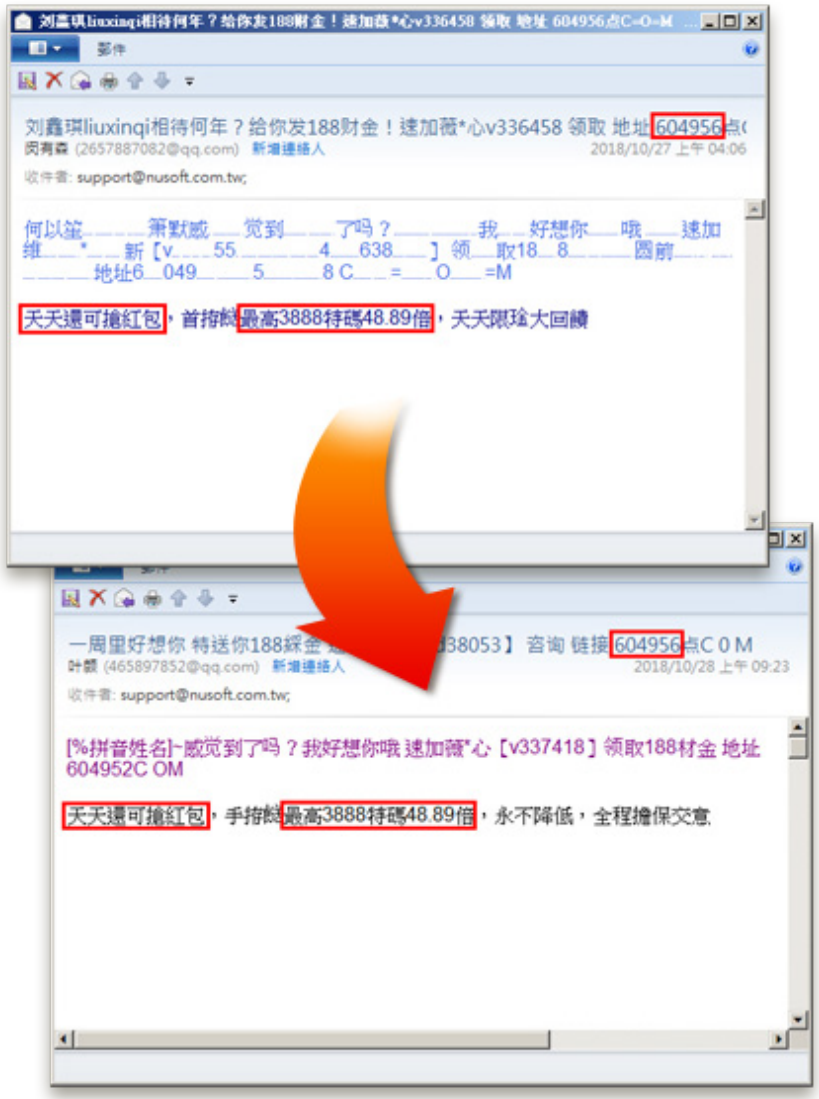
2. 會出現不同語言、排版但內容相似的勒索信件：

這種信件為了讓被害人能看的懂勒索的內容，又不確定被害人慣用的語系為何，所以會同時釋出不同語言但內容相似的版本。因這種信件使用文字完全不同，所以對於以往過濾系統來說是封全新信件。

AI Anti-Spam 會將信件分詞後，與所學 ham 和 spam 字詞資料庫比對。原本學習英文版勒索 spam，後續收到其他語系的類似信件，其內文字詞幾乎都沒學習過，故比對到的關鍵字詞權重值，就是判斷此類信件為 spam 的重要依據。

3. 常換句話說，改變排版方式的推銷郵件：

廣告信為了躲避過濾系統的防堵，會常把一些慣用詞以換句話說的方式呈現、改變排版方式或乾脆以圖片取代文字方式，讓過濾系統看不懂而放它一馬。



AI Anti-Spam 會將信件分詞後，與所學 ham 和 spam 字詞資料庫比對。廣告 spam 經學習後，後續收到變更部份內容、排版的類似信件，其內文、主旨的多個關鍵字串權重值，是判斷此類信件為 spam 的重要依據。

有鑑於此，新軟系統推出 AI Anti-Spam 機制：

a. 收集客戶日常往來的信件（透過舊有郵件過濾機制標示為 ham、spam），以 Word2Vec、Jieba 進行郵件內容分詞並轉換為有意義的大數據。

b. 採用 Google 的 TensorFlow 深度學習專案，一開始深度學習資料庫是空的，要將上述大數據投入 TensorFlow 的深度神經網路中進行預測，看預測的結果是否符合實際郵件屬性，這個過程就叫做建模；若預測的結果與實際郵件屬性不符，TensorFlow 就會採用梯度下降演算法，自動調整其內建參數，然後重複學習同一封郵件，直到預測結果與實際郵件屬性相符。在學習大量信件後，即可建構客戶端專屬的郵件預測模型。

c. 結合誤判回報功能，使用者僅要告知設備“那封信件判斷有誤”，系統會依使用者回報的信件屬性重新學習，來調整預測模型，讓 spam 辨識能更符合客戶端的電子郵件環境。

TensorFlow 深度神經網路通常會有數個階層，每個階層中會有多個神經元（由權重值、偏差值、輸入的變數所組成），在進行郵件學習時，採用的梯度下降演算法，一開始會先以亂數指定權重與偏差值，由於我們知道每筆郵件實際屬性，所以透過不斷修改權重與偏差值，讓最後預測結果與真正的答案相符（取得各神經元最佳的權重和偏差值組合）。

AI Anti-Spam 對於所有 spam 能更加精確的判斷，針對舊有郵件過濾機制無法處理的變種 spam，AI Anti-Spam 若學習過相關郵件，會分析變種 spam 中是否有相符的關鍵字詞、連結、表頭資訊，以有效阻擋；同時 AI Anti-Spam 會再學習辨識到的 spam，將整封信的內容新增為類似信件的評判依據；由此，AI Anti-Spam 可以不斷自我學習，增強各類 spam 辨識的靈敏度。

所以，舊有郵件過濾機制雖已達到 97% spam 辨識率，但 AI Anti-Spam 可以讓 spam 辨識率提升到 99.5% 以上；如果每天有 1000 封 spam 寄來，以往使用者會收到 30 封 spam，採用 AI Anti-Spam 則可讓收到的 spam 降到 5 封以下。

舉例來說，一封英文 spam 經 AI Anti-Spam 機制訓練後，相關郵件字詞數據會記錄在 spam 訓練資料庫中，若有變種 spam(假設是日文信) 則會比對其中有被訓練過的關鍵字詞權重，來有效判別此類信件。然後，會同時將整封變種日文 spam 進行訓練，往後有符合其中相關日文字詞的變種信，也可以準確被辨識，達到智慧學習的效果。

--- 全文完