Feedback of Speech III

本次的主題是深度學習在網路安全的應用，主要分成三個小主題來探討:

1. 何為網路安全與謠言散播問題:

本小節介紹了很多人類難以在第一時間分辨出來的釣魚網站，以及 Line 平台上的釣魚連結，而隨著詐騙網站的製作工具越來多也越來越方便，上述的詐騙事件可以說是層出不窮，而且這些跳於連結或是釣魚網站大多都是跟當時最熱門的議題掛勾，很容易吸引到許多受害者，但大多數的機器學習模型是需要歷史資料的來作為訓練資料的，因此要偵測近期熱門話題的詐騙事件就對困難，此外由於不同地區的用語不同，因此要偵測不同地區的詐騙事件就很困難。

2. 深度學習應用於網路安全的挑戰

直接將各式各樣的詐騙事件的內容具象化成機器學習模型可以解析是一個相當困難的挑戰，所以有時候會從詐騙者的角度思考會有哪些詐騙手段，而詐騙的手段會以廣灑釣餌為主，通常都會建立一個可拋棄式的帳號，用來收割詐騙得來的資訊。且詐騙者大多都是高度分工的團隊，甚至會採取分包的方式，而高度分工的方式就會產生固定的詐騙手段，而這就會令詐騙手段開始有一些模式可以偵測。例如強調”免費”，或是”趕緊”，”乾淨”... 等短期字眼就很常出現，像這類有比較短且有一些模式可以偵測的特徵，就可以丟掉常見的模型中進行二元偵測。

通常直接提示用戶遇到詐騙事件的話，用戶的不見得會直接採信模型的判斷結果，而如果在偵測到詐騙事件的同時提供”詐騙意圖”，以及”判斷依據”...等資訊，也就是模型的解釋性，用戶對模型的判斷結果的信賴程度就會大大增加，因為常見的詐騙手段會將詐騙的真實意圖藏在一大段文字訊息中，所以如果不提示用戶這個詐騙事件的真實意圖的話，用戶對模型的信賴程度會下降，而通常這可以透過文章分析的方式來挖出對方的真實意圖。

而模型常見的偵測議題如下： 透過 Bi-directional LSTM + Attention 的對 email 進行詐騙與否的二元分類，以及詐騙類型的多元分類，再來就是透過 CNN 模型對語句進行情緒分析，用以偵測詐騙與否，很多詐騙資訊是以短訊息的方式呈現，但是大多數模型都是透過長篇文章的語料庫來進行訓練的如 wiki 等，但是短訊息有很多模型無法辨認的特徵，如詞語縮寫，甚至是因為語句太短無法擷取足夠的特徵。

一些模型如 Bert 的偵測效果相當好，但是卻無法很好的發布在客戶端，因為除了偵測效果以外，模型運行條件也是考量點之一

3. 其他議題

除了比較肥大的深度學習模型以外，也可以透過 rumor busting database 如 TFC, MyGopen, Cofacts; 以及 semantic matching skill 如：universal sentence encoder + Approximate Nearest Neighbor 來進行謠言的偵測。此外，也可以將透過偵測結果誘導用戶到 rumor busting database 進行確認，如此一來這些偵測工具與 rumor busting database 可以達到互相成長的效果。

心得:

這次的演講讓我學到詐騙事件的分析方式，而且模型的準確率也不是唯一考量，因為模型運作條件可能並不符合實際環境的運作條件，或者說實務上，模型的訓練的前置資料處理動作，模型的訓練時長，或者是模型的滾動功能，模型判別時能夠給予哪些資訊...等都是很重要的考量點，從而造成一些簡單的，傳統的模型在使用上反而會比這些複雜的模型更重要。