國立清華大學電機工程學系

實作專題競賽成果摘要

# 語言與自然語言多模態偵謊辨識

# Multimodal Deception Detection by Language and Natural Language


組別：B270

領域：系統領域

指導教授：李祈均教授

組員：王瑭毅、冉宜軒

# 報告摘要

現今人工智慧技術發展日新月異，人類所觀察不出的事物可以透過人工智慧協助辨別，目前已有相當多研究展現了證實人工智慧具有預測或估計人類的外在表現和內部反應的能力。

在法務部廉政署偵訊過程中，多數被認定為嫌疑人的受詢人時常以謊言扭曲關鍵事實，以躲避追緝，若是能判斷出受詢人哪些言論有較高的說謊可能性，則可以使廉政官將對談內容聚焦，減少偵訊時間，以加速辦案效率。本作利用偵訊時的「音訊」與「文本」紀錄來進行偵謊，透過人工智慧模型來判斷受詢人的口供內容是否屬實。

我們從廉政官與受詢人每一次的問答之中萃取出文本特徵（BERT）、聲學特徵（Wav2Vec, ACO）以及行為特徵（CTD）作為辨別受詢人說謊的依據，根據我們的實驗結果，若只使用單一一種特徵，則模型的表現通常不佳，故我們將多種特徵融合，例如：文本特徵搭配聲學特或是聲學特徵搭配行為特徵等，可使模型獲得更完整的資訊，以利判斷。

此外，我們發現要偵測受詢人的說謊跡像，不單單只依靠受詢人的種種特徵，而是要將廉政官的反應也納入考量。當人類在說謊時，由於缺乏完整事件經歷，常將事件的細節部分模糊帶過，故可能引起對話者不自覺提高發問頻率，而此種行為被我們在行為特徵中發現，意味著受詢人的說謊行為可能顯露在廉政官的行為上，所以廉政官與受詢人的資訊對我們來皆是判別受詢人是否有隱瞞事實行為的重要依據。

由於廉政官在問詢時，會依循特定模式，會將整體談話內容分為數個主題段落，而每個主題段落的目的皆是試圖了解受詢人是否曾經執行過某個不法行為，一個段落大約會有數十次的答辯，而我們成功利用此種模式設計出多實例學習模型，來判別受詢人在每個主題段落中是否有意圖說謊的跡象，並同時提高逐句偵謊的表現，模型在針對主題段落層級之偵測結果 F1 score 最高可達 0.73，而逐句之偵測結果 F1 score 最高可達 0.72。

**關鍵字：**自然語言處理、行為訊號、多模態、多實例學習

# 報告內容

## 一. 研究背景

由於機器學習之技術日新月異，將其應用在各個領域已成為當今之趨勢，舉凡電腦影像、音訊、多媒體等皆廣泛應用機器學習之技術，而在隨著科技的進步，測量並收集來自人類的訊號資訊逐漸變得普及，有無數種方式可以用來解讀這資訊，藉此量化或計算來自人類的訊號，例如情感運算[1]、社會訊號處理[2]、行為訊號處理[3]。針對行為訊號的部分，已有相當多研究展現了如何利用行為訊號來預測或估計人類的外在表現抑或內部反應。

雖然欺騙行為是人類生活的自然組成部分，但眾所周知，人類並不善於體悟到他人的說謊行為，對於此周惶振學長、劉奕汶教授以及李祈均教授曾透過計算聲學、文本訊息、對話時間行為設計自動偵謊模型[4]，並發現從詢問者的對話行為可以觀察到對答者的欺騙行為以及一些聲學特徵與文本特徵是檢測謊言的有效指標。故利用機器學習手法，能有效解讀人類訊號已成為不爭的事實，因此，在本作中我們將利用機器學習的方式，來協助我們判讀人類訊號。

## 二. 研究目的

法務部廉政署成立的其中一項目的為打擊台灣公務員貪污以及瀆職，而針對違法行為的公務員，在經由調查組查證有貪污瀆職之事實後，將由檢察官進行起訴，而在起訴過程中可能缺乏決定性證據，而導致無法將嫌疑人定罪，此時，將傳喚嫌疑人至廉政署與廉政官進行對談，廉政官並非如同電影情節是虎背熊腰之警方人員擔任，而通常出任之人皆性情和藹面露慈善，期望透過在與嫌疑人閒話家常時，使嫌疑人無意間透露出關鍵案情，以供檢察官在法庭上能以此作為證據，請求法官將嫌疑人定罪。不過，由於偵訊階段嫌疑人皆抱持著高度戒心，因此大多數情形嫌疑人總是能以謊言扭曲關鍵事實，若是能判斷出嫌疑人哪些言論有較高的說謊可能性，則可以使廉政官將對談內容聚焦，減少偵訊時間，以加速辦案效率。

因此，在廉政署對於欺瞞行為的辨識需求下，我們藉由聲學、文本以及對話行為來量化受詢人的逐句答辯狀況，並設計自動偵謊系統，辨別受詢人的回應是否屬實。

---

[1] Picard, R. W. (1995). Affective computing.

[2] Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., &Schroeder, M. (2012). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. IEEE Transactions on Affective Computing, 3(1), 69-87.

[3] Narayanan, S., & Georgiou, P. G. (2013). Behavioral signal processing: Deriving human behavioral informatics from speech and language. Proceedings of the IEEE, 101(5), 1203-1233.

[4] Huang-Cheng Chou, Yi-Wen Liu, Chi-Chun Lee. Automatic Deception Detection using Multiple Speech and Language Communicative Descriptors in Dialogs. Cambridge University Press in association with Asia Pacific Signal and Information Processing Association, vil. 10, e5, 1-9.

## 三. 研究方法

### 3.1 資料集與資料前處理

本作所使用之資料皆由法務部方面專業人士標記完成，此資料集一共收錄 40 位廉政官與受詢人的對話紀錄，在經過篩選汰除掉有瑕疵或是不適當的案例後，最終剩下 33 則案例的對話紀錄，而每筆案例的對話時長不等，一共剪輯成 13265 筆的音檔，每一筆音檔又可分成廉政官的提問與受詢人的答辯部分，其中被標註為真實的音檔約佔 87%，換而言之，被標註為謊言的音檔約佔 13%。

| 修正前 | 40位受詢人 |
|---|---|
| 基準（0） | 1746 |
| 真實（1） | 13276 |
| 謊言（2） | 2548 |
| 總和 | 17572 |

| 修正後 | 33位受詢人 |
|---|---|
| 真實（0） | 11558 |
| 謊言（1） | 1707 |
| 總和 | 13265 |

Table 1 - dataset

### 3.2 特徵萃取

本作一共萃取以下四種特徵：

1) Textual Embedding Feature (BERT)：768 維的文本特徵。
2) wav2vec 2.0 Feature (Wav2Vec)：512 維的聲學特徵。
3) Acoustic-prosodic Feature (ACO)：988 維的聲學特徵。
4) Turn-Level Conversational Temporal Dynamics Feature (CTD)：7 維的行為特徵。

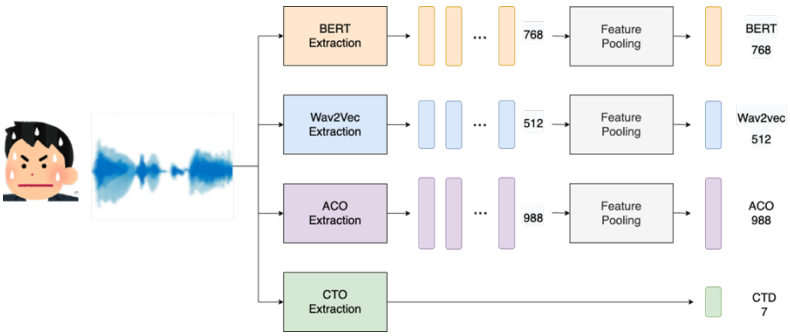


Figure 1 - illustration of feature extraction

### 3.3 人工智慧模型

本作一共使用以下三種人工智慧模型：

1) 長短記憶模型（Long Short-term Memory, LSTM）[5]
2) Transformer[6]
3) Hopfield Layer[7]

### 3.4 多實例學習（Multi-Instance Learning, MIL）

依據廉政官的提問模式，相鄰的數十句的問答會圍繞著同一個不法事實，因此我們在逐句判斷受詢人是否說謊之餘，利用多實例學習的方式依據數十句對話的模型輸出結果加以判斷受詢人是否對該不法事實有欺瞞之嫌。

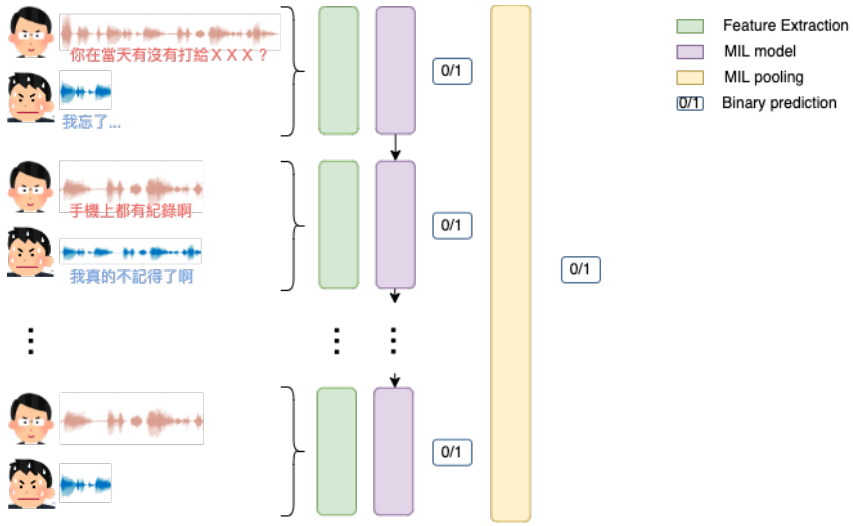


Figure 2 - illustration of multi-instance learning

[5] Sepp Hochreiter, Jürgen Schmidhuber, Long Short-term Memory, (1997) Neural Computation, 9(8), 1735-1780.
[6] Ashish Vaswani, Noam Shazeer et la., Attention Is All You Need, 31st Conference on Neural Information Processing Systems (NIPS 2017)
[7] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner et al. (2021) Hopfield Networks is All You Need.

# 四. 研究結果

## 4.1 實驗一（Single Modality Training）

本實驗架構針對使用單一特徵（single modality feature）設計，將同一案例之全部資料輸入模型中，比較單獨使用受詢人答辯（monologue）與同時使用廉政官提問與受詢人答辯（dialogue）時各種模型的表現。

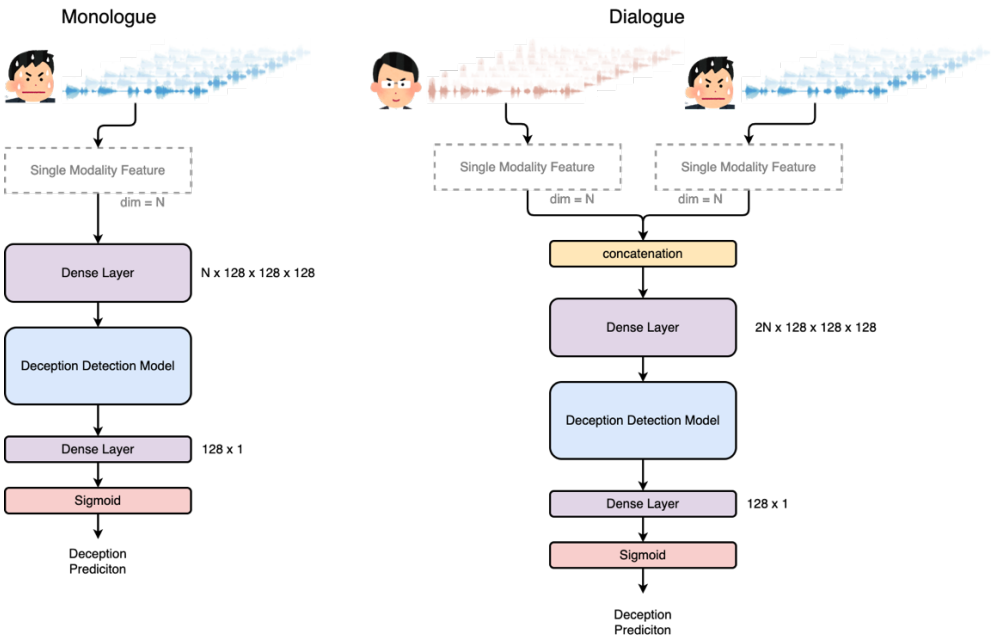


Figure 3 - illustration of the model framework of experiment I

| parameter | learning rate | epoch | batch size | wighted sampling | loss function |
|---|---|---|---|---|---|
| value | 0.001 | 30 | 64 | TRUE | cross-entropy |

Table 2 - parameter of experiment I

| model | feature | feature pooling | conversation | ACC | UAR | Precision | F1 score |
|---|---|---|---|---|---|---|---|
| LSTM | BERT | mean | monologue | 0.7669 | 0.5576 | 0.5515 | 0.5539 |
| LSTM | BERT | mean | dialogue | 0.7337 | 0.5755 | 0.5535 | 0.5566 |
| LSTM | Wav2Vec | mean | monologue | 0.6936 | 0.5532 | 0.5302 | 0.5215 |
| LSTM | Wav2Vec | mean | dialogue | 0.6848 | 0.5358 | 0.5202 | 0.5089 |
| LSTM | ACO | mean | monologue | 0.8467 | 0.4954 | 0.4821 | 0.4762 |
| LSTM | ACO | mean | dialogue | 0.8097 | 0.5439 | 0.5507 | 0.5466 |
| LSTM | CTD | mean | monologue | 0.7582 | 0.5425 | 0.5367 | 0.5386 |
| LSTM | CTD | mean | dialogue | 0.7506 | 0.5402 | 0.5333 | 0.5350 |
| Transformer | BERT | mean | monologue | 0.8180 | 0.5585 | 0.5845 | 0.5658 |
| Transformer | BERT | mean | dialogue | 0.8308 | 0.6335 | 0.6474 | 0.6397 |
| Transformer | Wav2Vec | mean | monologue | 0.7803 | 0.5298 | 0.5277 | 0.5286 |
| Transformer | Wav2Vec | mean | dialogue | 0.7338 | 0.5228 | 0.5160 | 0.5142 |
| Transformer | CTD | mean | monologue | 0.7860 | 0.5725 | 0.5686 | 0.5704 |
| Transformer | CTD | mean | dialogue | 0.7214 | 0.5974 | 0.5617 | 0.5633 |

Table 3 - result of experiment I

## 4.2 實驗二（Multiple Modality Training）

有鑒於實驗一使用單一特徵訓練模型時，所得到表現極為有限，是故本實驗使用多項特徵作為訓練資料並使用晚期融合（late fusion）的手法將相異特徵融合，作為偵謊模型的輸入。此外，本實驗共可分為兩部分，第一部分使用相異的特徵池化，挑選出最合適的池化方式。

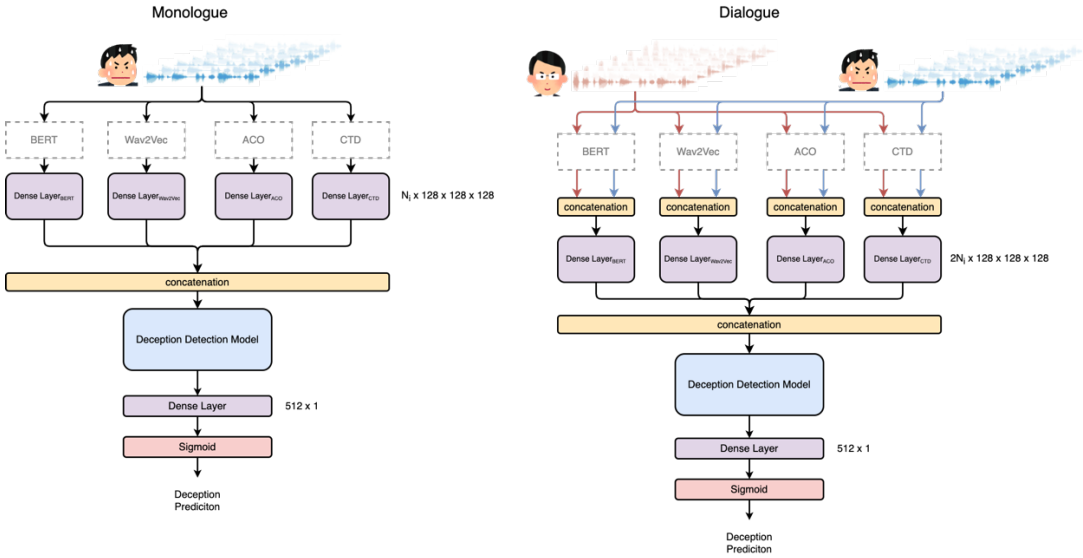


Figure 4 - illustration of model framework of experiment II

| parameter | learning rate | epoch | batch size | wighted sampling | loss function |
|---|---|---|---|---|---|
| value | 0.001 | 30 | 64 | TRUE | cross-entropy |

Table 4 - parameter of experiment II-1

| model | features | feature pooling | conversation | random seed | ACC | UAR | Precision | F1 scroe |
|---|---|---|---|---|---|---|---|---|
| LSTM | BERT, Wav2Vec, CTD | min | monologue | 2022 | 0.6515 | 0.6369 | 0.5652 | 0.5398 |
| LSTM | BERT, Wav2Vec, CTD | min | monologue | 6 | 0.5656 | 0.5781 | 0.5353 | 0.4765 |
| LSTM | BERT, Wav2Vec, CTD | min | monologue | 25 | 0.7378 | 0.5493 | 0.5333 | 0.5335 |
| LSTM | BERT, Wav2Vec, CTD | min | monologue | 9 | 0.7336 | 0.5926 | 0.5567 | 0.5583 |
| LSTM | BERT, Wav2Vec, CTD | min | monologue | 14 | 0.4891 | 0.6124 | 0.5518 | 0.4426 |
| LSTM | BERT, Wav2Vec, CTD | mean | monologue | 2022 | 0.7782 | 0.5762 | 0.5601 | 0.5652 |
| LSTM | BERT, Wav2Vec, CTD | mean | monologue | 6 | 0.7758 | 0.5628 | 0.5507 | 0.5545 |
| LSTM | BERT, Wav2Vec, CTD | mean | monologue | 25 | 0.7949 | 0.5798 | 0.5697 | 0.5738 |
| LSTM | BERT, Wav2Vec, CTD | mean | monologue | 9 | 0.8035 | 0.5560 | 0.5572 | 0.5566 |
| LSTM | BERT, Wav2Vec, CTD | mean | monologue | 14 | 0.8197 | 0.5266 | 0.5381 | 0.5287 |
| LSTM | BERT, Wav2Vec, CTD | max | monologue | 2022 | 0.5814 | 0.5102 | 0.5048 | 0.4596 |
| LSTM | BERT, Wav2Vec, CTD | max | monologue | 6 | 0.6133 | 0.6140 | 0.5526 | 0.5123 |
| LSTM | BERT, Wav2Vec, CTD | max | monologue | 25 | 0.5682 | 0.6321 | 0.5592 | 0.4938 |
| LSTM | BERT, Wav2Vec, CTD | max | monologue | 9 | 0.5552 | 0.5859 | 0.5386 | 0.4735 |
| LSTM | BERT, Wav2Vec, CTD | max | monologue | 14 | 0.6048 | 0.6318 | 0.5600 | 0.5137 |
| Transformer | BERT, Wav2Vec, CTD | min | monologue | 2022 | 0.7152 | 0.5873 | 0.5505 | 0.5478 |
| Transformer | BERT, Wav2Vec, CTD | min | monologue | 6 | 0.6843 | 0.6327 | 0.5671 | 0.5547 |
| Transformer | BERT, Wav2Vec, CTD | min | monologue | 25 | 0.7037 | 0.5782 | 0.5441 | 0.5384 |
| Transformer | BERT, Wav2Vec, CTD | min | monologue | 9 | 0.6286 | 0.6205 | 0.5563 | 0.5223 |
| Transformer | BERT, Wav2Vec, CTD | min | monologue | 14 | 0.7177 | 0.6281 | 0.5706 | 0.5692 |
| Transformer | BERT, Wav2Vec, CTD | mean | monologue | 2022 | 0.7964 | 0.5454 | 0.5455 | 0.5454 |
| Transformer | BERT, Wav2Vec, CTD | mean | monologue | 6 | 0.8095 | 0.5377 | 0.5445 | 0.5402 |
| Transformer | BERT, Wav2Vec, CTD | mean | monologue | 25 | 0.8349 | 0.5653 | 0.5906 | 0.5735 |
| Transformer | BERT, Wav2Vec, CTD | mean | monologue | 9 | 0.8008 | 0.6019 | 0.5865 | 0.5926 |
| Transformer | BERT, Wav2Vec, CTD | mean | monologue | 14 | 0.8066 | 0.5878 | 0.5816 | 0.5844 |
| Transformer | BERT, Wav2Vec, CTD | max | monologue | 2022 | 0.7551 | 0.5749 | 0.5522 | 0.5561 |
| Transformer | BERT, Wav2Vec, CTD | max | monologue | 6 | 0.7501 | 0.6175 | 0.5734 | 0.5793 |
| Transformer | BERT, Wav2Vec, CTD | max | monologue | 25 | 0.7691 | 0.6322 | 0.5871 | 0.5967 |
| Transformer | BERT, Wav2Vec, CTD | max | monologue | 9 | 0.7470 | 0.6107 | 0.5692 | 0.5742 |
| Transformer | BERT, Wav2Vec, CTD | max | monologue | 14 | 0.7145 | 0.6566 | 0.5827 | 0.5801 |

Table 5 - result of experiment II-1

而第二部分我們嘗試使用相異的特徵組合，來確認加入廉政官的提問資訊是否有助於模型偵測說謊的動作。

| parameter | learning rate | epoch | batch size | wighted sampling | loss function |
|---|---|---|---|---|---|
| value | 0.001 | 30 | 64 | TRUE | cross-entropy |

Table 6 - parameter of experiment II-2

| model | features | feature pooling | conversation | ACC | UAR | Precision | F1 score |
|---|---|---|---|---|---|---|---|
| LSTM | BERT, Wav2Vec | mean | monologue | 0.7826 | 0.5431 | 0.5390 | 0.5406 |
| LSTM | BERT, Wav2Vec, CTD | mean | monologue | 0.7826 | 0.5431 | 0.5390 | 0.5406 |
| LSTM | BERT, ACO | mean | monologue | 0.7911 | 0.5459 | 0.5440 | 0.5449 |
| LSTM | BERT, ACO, CTD | mean | monologue | 0.7911 | 0.5459 | 0.5440 | 0.5449 |
| LSTM | BERT, CTD | mean | monologue | 0.7669 | 0.5576 | 0.5515 | 0.5539 |
| Transformer | Wav2Vec, CTD | mean | monologue | 0.8186 | 0.5654 | 0.5746 | 0.5692 |
| Transformer | BERT, CTD | mean | monologue | 0.8297 | 0.5653 | 0.6017 | 0.5750 |
| Transformer | BERT, ACO, CTD | mean | monologue | 0.8476 | 0.5648 | 0.6103 | 0.5763 |
| Transformer | BERT, Wav2Vec, ACO | mean | monologue | 0.8287 | 0.5697 | 0.5875 | 0.5763 |
| Transformer | BERT, Wav2Vec, CTD | mean | monologue | 0.8424 | 0.5673 | 0.6030 | 0.5776 |
| Transformer | Wav2Vec, ACO, CTD | mean | monologue | 0.8313 | 0.5720 | 0.5921 | 0.5794 |
| Transformer | BERT, Wav2Vec | mean | monologue | 0.8432 | 0.5692 | 0.6060 | 0.5800 |
| Transformer | BERT, ACO | mean | monologue | 0.8281 | 0.5781 | 0.5930 | 0.5841 |
| Transformer | BERT, Wav2Vec, ACO, CTD | mean | monologue | 0.8585 | 0.5699 | 0.6408 | 0.5850 |
| LSTM | ACO, CTD | mean | dialogue | 0.8097 | 0.5439 | 0.5507 | 0.5466 |
| LSTM | BERT, CTD | mean | dialogue | 0.7337 | 0.5755 | 0.5535 | 0.5566 |
| LSTM | BERT, Wav2Vec, ACO | mean | dialogue | 0.7770 | 0.5700 | 0.5560 | 0.5605 |
| LSTM | BERT, Wav2Vec, ACO, CTD | mean | dialogue | 0.7770 | 0.5700 | 0.5560 | 0.5605 |
| LSTM | BERT, ACO | mean | dialogue | 0.7585 | 0.5844 | 0.5587 | 0.5636 |
| LSTM | BERT, ACO, CTD | mean | dialogue | 0.7585 | 0.5844 | 0.5587 | 0.5636 |
| LSTM | BERT, Wav2Vec | mean | dialogue | 0.7835 | 0.5759 | 0.5621 | 0.5669 |
| LSTM | BERT, Wav2Vec, CTD | mean | dialogue | 0.7835 | 0.5759 | 0.5621 | 0.5669 |
| Transformer | BERT, Wav2Vec, CTD | mean | dialogue | 0.8488 | 0.5740 | 0.6198 | 0.5868 |
| Transformer | BERT, Wav2Vec, ACO | mean | dialogue | 0.8422 | 0.5780 | 0.6113 | 0.5889 |
| Transformer | BERT, Wav2Vec | mean | dialogue | 0.8681 | 0.5796 | 0.6798 | 0.5995 |
| Transformer | ACO, CTD | mean | dialogue | 0.8382 | 0.5993 | 0.6191 | 0.6074 |
| Transformer | BERT, Wav2Vec, ACO, CTD | mean | dialogue | 0.8379 | 0.5997 | 0.6191 | 0.6077 |
| Transformer | BERT, ACO | mean | dialogue | 0.8189 | 0.6277 | 0.6138 | 0.6199 |
| Transformer | BERT, CTD | mean | dialogue | 0.8122 | 0.6259 | 0.6195 | 0.6225 |
| **Transformer** | **BERT, ACO, CTD** | **mean** | **dialogue** | **0.8390** | **0.6290** | **0.6358** | **0.6322** |

Table 7- result of experiment II-2

### 4.3 實驗三（Multi-Instance Based Training）

由於廉政官的提問方式將數十句對話圍繞在同一主題，我們想了解受詢人是否對於此主題意圖隱瞞事實，因此我們設定相異大小的主題段落（paragraph size），除了逐句判斷說謊的可能性外，也偵測受詢人在此範圍之中是否有說謊的跡象。

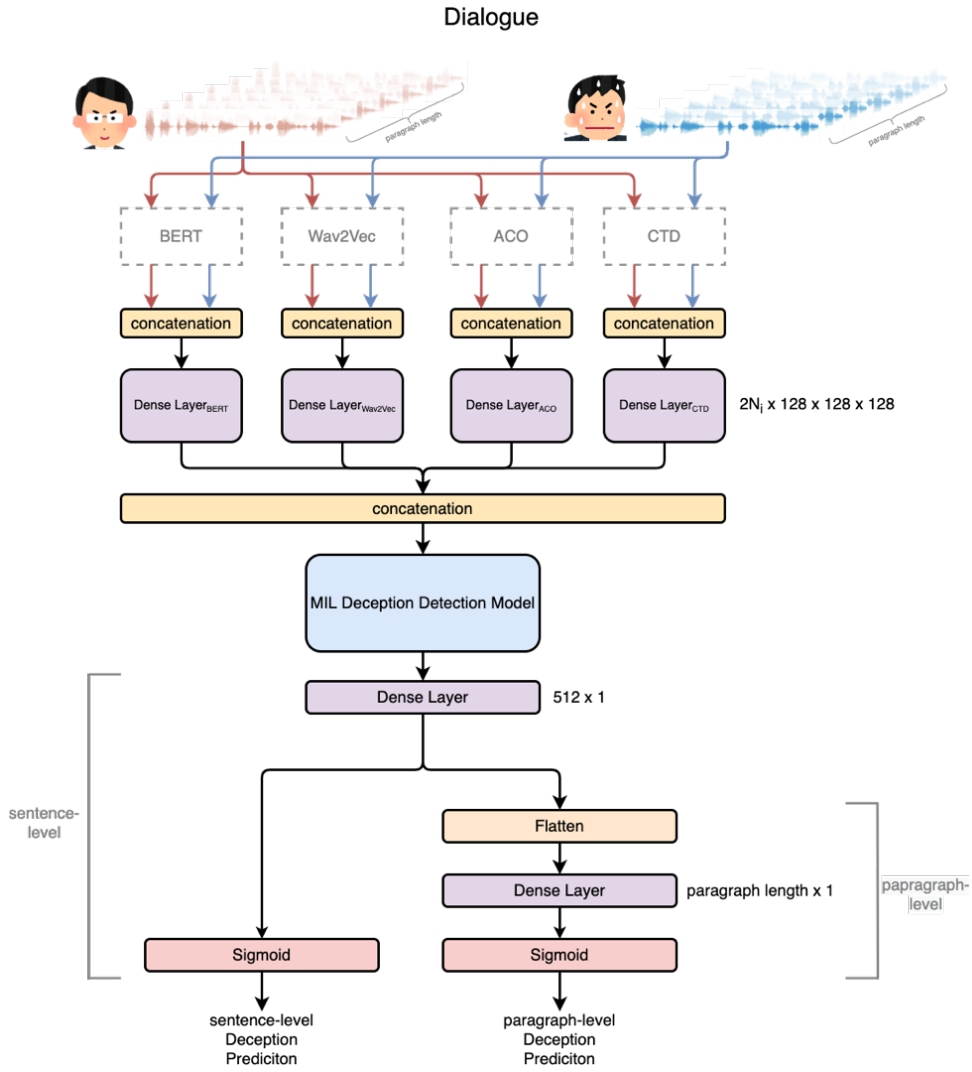


Figure 5 - illustration of model framework of experiment III

| parameter | learning rate | epoch | batch size | wighted sampling | early-stop | loss function |
|---|---|---|---|---|---|---|
| value | 0.001 | 50 | 64 | TRUE | TRUE | cross-entropy |

Table 3 - parameter of experiment III

| | | | | | sentense-level | | | | paragraph-level | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| features | model | feature pooling | paragraph size | conversation | ACC | UAR | Precision | F1 score | ACC | UAR | Precision | F1 score |
| BERT, ACO, CTD | Hopfield | mean | 5 | dialogue | 0.7916 | 0.7480 | 0.6457 | 0.6655 | 0.5533 | 0.6448 | 0.5720 | 0.4991 |
| BERT, ACO, CTD | Hopfield | mean | 10 | dialogue | 0.8163 | 0.7747 | 0.6701 | 0.6958 | 0.8203 | 0.7450 | 0.6942 | 0.7127 |
| **BERT, ACO, CTD** | **Hopfield** | **mean** | **15** | **dialogue** | **0.8810** | **0.7163** | **0.7357** | **0.7253** | **0.8633** | **0.7072** | **0.7775** | **0.7332** |
| BERT, ACO, CTD | Hopfield | mean | 20 | dialogue | 0.8294 | 0.7839 | 0.6810 | 0.7089 | 0.6390 | 0.6237 | 0.5820 | 0.5696 |

Table 9 - result of experiment III

## 五. 結論

在這一年內調配各種模型配對不同方式來訓練，由淺入深地觀察與統整各模型訓練結果，得到以下結論：

1. 使用多種特徵進行訓練可以提供模型更多資訊，相較於使用單一特徵時的 F1 score 總是在 0.5 附近左右徘徊，唯獨使用 BERT 時可以到 0.6，我們推測主因為人類說謊時，容易在句尾加上特定語助詞，而文本特徵較聲學特徵較易捕捉此種現象。
2. 使用對話（dialogue）類型的資料相較於僅使獨白（monologue）可以有效提升偵謊模型的表現，因為當受詢人意圖捏造謊言掩蓋事實時，容易使前後文有交代不清的現象，而此時廉政官會不自覺地增加提問頻率，而此種現象被 CTD 特徵所捕捉，所以廉政官的行為表現也透露了受詢人試圖說謊的行為。
3. 根據廉政官提問模式，我們所建立的多實例學習模型可以有效偵測受詢人在一主題段落的範圍內是否有說謊的跡象，其中主題段落的長度約為 15 句話，同時，也提升針對逐句偵謊的表現，最佳的模型在主題段落層級的偵謊結果 F1 score 可以達到 0.73，而逐句層級的偵謊結果 F1 score 可以達到 0.72。

# 心得感想

經由這次的專題，我們了解到凡是需要機器學習參與的領域，都需要抽象概念轉換以轉變電腦可處理的數字，例如音頻、文字的特徵使用，也對於 BERT 和 Wav2Vec 2.0 有初步的了解，我們一探到機器學習應用涉及了很多的數學運算，不管是底層的 self-attention 到後期的訓練句數關聯等等，都和數學運算環環相扣，不禁感嘆生活中，許多利用數學描述的事物看似極為抽象，但卻又極為真實，很多需要經過 AI 判斷的小事情，例如現今熱門的 GPS 導航、AI 繪圖等等，背後也是經過千萬條數學式子、模型輸出入、判斷等等而來，並且也都是前人智慧所累積，令人不禁驚嘆其背後所蘊含的邏輯架構是如何被發想出來的，因此，感謝電機系李祈均老師和彼得，給予我們這次機會讓我們接觸人工智慧模型的發想過程，彼得甚至還帶我們到廉政署裡去一探偵詢時的場景，讓我們對於此計畫整個脈絡有深刻的理解，最終也要感謝 BIIC 的各位學長姐，讓我們在實驗室有好笑一又有點奇怪的回憶。