國立清華大學 電機工程學系

實作專題研究成果報告

Deception Detection System

測謊系統技術

專題領域： 通訊所

組　　別： B167

指導教授：李祈均

組員姓名：李秉澤

研究期間:110年3月1日至110年12月底止，計9個月

Abstract

A generally difficult task for humans is to detect whether someone is telling the truth for a long time and this happens almost everyday in the court system specifically .
In the situation, one tool that is occasionlly used is a polygraph test in the situation, for measuring whether someone is being intentionally deceptive. Unfortunately , if we use such a polygraph test is difficult to set up and requires physical touch to get working, and even then it may not achieve fully accuracy .
Especially in a court setting , therefore other methods may be necessary to attempt to identify deception. A deception dataset is provided by P´erez-Rosas et al. that covers the use of deception in the court settings . This public dataset was constructed by taking recordings of public testimonies of witnesses and the defendants, classifying what they said as truthful or deceptive based on the final result of the court case. So, this project is going to expend and analyze deeply this database

摘要

長久以來，對人類來說，一個普遍困難的任務是檢測某人是否在說真話，特別是在法院系統中幾乎每天都發生，而在那個情況下檢驗測謊的最佳工具便是使用的一個測謊儀。
用於測量某人是否在故意欺騙。不幸的是，若是我們要使用這種測謊儀是很難設置的，而且需要身體接觸才能工作，即使如此，它也時常無法達到完全準確度[1]。因此，特別是在法庭上我們可能需要其他方法來試圖識別欺騙行為。

P´erez-Rosas 等人提供了一個欺騙資料集，涵蓋了在法庭上使用的欺騙在法庭上的使用[2]。這個公開的資料集是通過對證人和被告的公開證詞進行錄音來構建的，為證人和被告的公開證詞，並將他們所說的話歸類為真實或欺騙性的基於法庭案件的最終結果，而此專題便是將此數據庫加以擴大與深度地分析。

目錄

一、前言

此專題進行了一個測謊數據特徵實驗，測試了各種模型和特徵交叉比對後的結果，並突出了 logistic regression 在該資料庫的行為特徵分類中有最佳效果。此外，由實驗結果也得到解析特徵可能在某些組合中提供幫助，但它並不總是能幫助模型識別欺騙，許多的測謊相關的論文也有提及需要注意的是在不同類別的解析特徵中看到的差異可能只是因為該數據量有限，所以在處理數據的部分有多花點時間去理解和整合。

至於我對此專題的想法便是基於對一般對運行欺騙檢測有意見，特別是在法庭環境中，測謊儀測試若是在不夠足量的數據分析下在很多情況下是其實是不准確的[1]，而在作者所做的最大的假設是，法庭案件中對雙方的回答的作出的結果是完全的仲裁者，但是這種情況只有在證人席上的人不是一個好的說謊者時才行得通。而此專題便是將主要問題建立於是在這個數據集的建構分析，並且希望此專題可以建構出一個數量更大且分析透徹的數據集，便可以在法院判決時輔助法院在 數據不足以確認指控時審判的依據，而這也是目前便是有許多人致力於建構測謊數據集的建構與分析[7]。

二、原理分析與系統設計

1. 原理分析

此專題進行了一個特徵集透過不同 model 的比較實驗，包括在論文中建議的特徵以及 PDTB (Penn Discourse Treebank，賓州篇章樹庫)[3]和 RST(Rhetorical Structure Theory Discourse Treebank，修辭結構理論樹庫) [4]中特殊提取的一些特徵，並且使用 RST 和 PDTB 的原本的解析器來解析數據集中的語句[5]-[6]，視覺檢查結果便標記後顯示有些句子無法被有效解析，而至於所有成功的 PDTB 解析都是從一個明確的標籤開始的，這些部分在論文中有提及可能原因為使用的是舊的數據分析器，並且在檢驗中可能有需許多尚為考量的點。

2. 系統設計

數據集 (Dataset):

the feature set experiment using the features described by P´erez-Rosas et al. of unigrams, bigrams, and behavioral features[2]

特徵 (Features):

OtherGestures, Smile, Laugh, Scowl, otherEyebrowMovement, Frown, Raise, OtherEyeMovements, Close-R, X-Open, Close-BE, gazeInterlocutor, gazeDown, gazeUp, otherGaze, gazeSide, openMouth, closeMouth, lipsDown, lipsUp, lipsRetracted, lipsProtruded, SideTurn, downR, sideTilt, backHead, otherHeadM, sideTurnR, sideTiltR, waggle, forwardHead, downRHead, singleHand, bothHands, otherHandM, complexHandM, sidewaysHand, downHands, upHands,

分類 (Classification):

deceptive and truthful

2.1 解析數庫特徵

從 RST parse trees 中，從每棵分析樹中提取依附特徵（ 句子與句子、句子與單字、單字與單字間彼此的相互關係 ）並排除涉及分析樹尾端的依附特徵（ 與實際依附關係過於遙遠因而忽略不計 ）以提高分析的準確率，然後我們計算每個依附特徵被使用的次數，並將這些計算成一個數量集並用作為一個特徵分析，最後整個數據集經過 RST 分析提取後便可得到 155 個特徵，並將最高的依附關係組合中最高的 5 組作出表格比較，同理我們將 pdtb parse trees 作相同的處理，並且一樣捨棄分析樹尾端的依附特徵，並且發現最高的依附關係組合中只有 4 個特徵，將實驗結果都作出表格後作比較。

實驗結果

表 1 (RST):

此表格顯示文字檔在 Truth label 和 Deception label 之中 count 最多的前 5 個依附特徵

Truth	Count
elaboration[n][s] → elaboration[n][s] elaboration[n][s]	19
same-unit[n][n] → elaboration[n][s] elaboration[n][s]	6
joint[n][n] → attribution[s][n] joint[n][n]	6
elaboration[n][s] → elaboration[n][s] joint[n][n]	6
elaboration[n][s] → attribution[s][n] elaboration[n][s]	6
Deception	**Count**
elaboration[n][s] → elaboration[n][s] elaboration[n][s]	12
same-unit[n][n] → elaboration[n][s] elaboration[n][s]	5
elaboration[n][s] → elaboration[n][s] attribution[s][n]	5
same-unit[n][n] → elaboration[n][s] attribution[s][n]	4
elaboration[n][s] → elaboration[n][s] explanation[n][s]	4

表 2 (PDTB):

此表格顯示文字檔在 Truth label 和 Deception 之中 count 最多的前 4 個依附特徵

Truth	Count
explicit=expansion	84
explicit=temporal	47
explicit=contingency	43
explicit=comparison	13
Deception	**Count**
explicit=expansion	88
explicit=contingency	49
explicit=temporal	30
explicit=comparison	27

結果分析:

這些特徵在 rst 分佈中有很明顯的部分為

elaboration[n][s] → elaboration[n][s] elaboration[n][s]

而這代表這不管是在說謊還是實話時,回答者總是用完整闡述語句,去作辯答的部分,這是相當合理的,因為受審的證人或被告通常是在解釋他們所說的話,這也與此資料庫收集的特定族群(受審的證人或被告)有關。

而這些特徵在 pdtb 中的分佈基本相同,唯一的區別是在 temporal 中真實陳述比在欺騙陳述中更常見,這代表著那些誠實說話的人有更好地了解事件何時發生,並且可以完整地敘述事件的發生過程,而欺騙者通常在敘述事件的確切時間之時會含糊不清的作出答覆。

我們可以依照此分析結果對於 rst 與 pdtb 的分類結果對原本的資料庫有更深的認識,也提供後續經過模型處理後的數據結果有合理的解釋。

2.2 數據庫經過 model 分析

將原本的Real-life_Deception_Dataset經過不同7處理方式（uniX, biX, triX,behavFeatures, rstX, pdtbX ）處理為7組不同的特徵集，並將特徵集彼此互相搭配形成一個龐大的特徵組合集(127 組)，接著將這些特徵組合集經過5組設計好的model (Logistic Regression, Support Vector Machines,and Decision Trees.Random Forest,K-nearest Neighbors)作分析，

並將各組結果作Kfold交叉驗證10次後取平均數，以提供較高的準確率，並且透過4種不同的效能衡量指標去評估此模型的準確與完整性,此實驗與先前作者在分析數據庫的不同為將原先的處理方式由 4 種提高為 7 種變化性更高的特徵集,並且將原本分析的model由3組提高為5組已透過更多樣的方式去分析此數據集，因此在Kfold交叉驗證後的實驗結果數據量由原先的372組 ((2^5-1) X 3 X 4) 提高為 2540 組((2^7-1) X 5 X 4)，有多了超過2000多組的實驗數據可以來評估所設計的model。

最開始會將文字的正規化（Text Normalisation）:

i. 小寫轉換（Lowercase Conversion）:

首先去除大小寫的差異，按照慣例將所有文字轉成小寫

ii.語幹提取（Stemming）

在語言學中，詞幹（word stem）表示一個單詞中最基本且核心的形式

例如:

friendships 由詞幹friendship與詞綴 -s所組成，

friendship 則是由 詞幹friend 與詞綴 -ship 所構成，

因此詞幹的提取基於不同理念或不同演算法，有時會得到不同的結果。

iii. 詞形還原（Lemmatisation）(By Porter Stemming Algorithm)

將所得到的詞幹與詞綴分開，並保留詞幹的的部分，然而萃取詞幹並未能完全滿足減少詞形變化（inflection）的需求，因此需要找尋更能代表單詞基本形式－詞位（lemma）

例如:

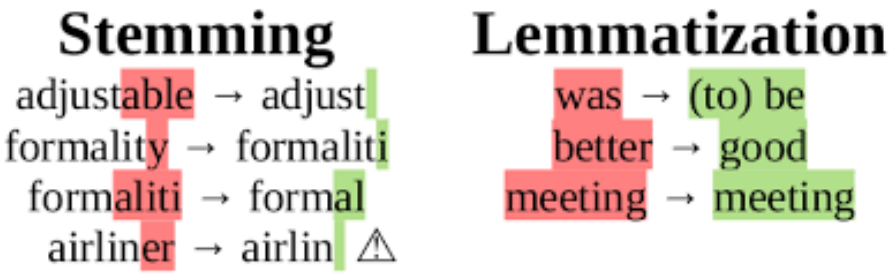
sings、 singing、 sang、 sung 共享同一個詞位 sing。

我選擇 NLTK.stem 模組中收錄的 WordNetLemmatizer（普林斯頓大學所建立的免費公開詞彙資料庫)類別找出詞位

iv. 停用詞去除（Stopword Removal）

在文句中有些單詞並對於詞義的傳達並無太大的作用，如a/ an、the、is/ are等，被稱之為停用詞（stop words），我們在此將她提取掉。

文字的正規化後結果:



圖片來源：DEVOPEDIA

model(5 個)

(1) KNeighborsClassifier (K-nearest Neighbors) ：找出 K 個跟測試資料最接近的點，再統計這些點的分類做為預測結果

(2) LogisticRegression (Logistic Regression)： 為一個二元分類器，藉由邏輯斯函數來將特徵資料投射到介於 0 到 1 之間的值，來判斷資料是否屬於某個分類

(3) LinearSVC (Linear Support vector machine)： SVM 為基於統計學習的監督式演算法，透過找出一個超平面，使之將兩個不同的集合分開的模型，而 LinearSVC 便是 SVM 透過調整超參數 C 來達 weight regularization 來限制模型的複雜度而形成的線性可分支持向量機

(4) DecisionTreeClassifier (Decision Tree)： 從最後上方的樹根開始將資料的特徵將資料分割到不同邊，根據獲得最大的資訊增益(Information gain)進行分工

(5) RandomForestClassifier (Random Forest)： 當作是多個決策樹組合而成的，並隨機從中選取決策樹作運算

特徵數據集(7 組):

(1) unigram (一元語法)： 將原本的句子分為單一單字的單位

例: Cys-Gly-Leu-Ser-Trp ------- Cys, Gly, Leu, Ser

(2) bigram (二元語法)： 將原本的句子分為兩個單字為一組的單位

例: Cys-Gly-Leu-Ser-Trp ------- Cys-Gly, Gly-Leu, Leu-Ser, Ser-Trp

(3) trigram (三元語法)： 將原本的句子分為三個單字為一組的單位

例: Cys-Gly-Leu-Ser-Trp ------- Cys-Gly-Leu, Gly-Leu-Ser, Leu-Ser-Trp

(4) behavFeatures = 原始 dataset 中的句子

(5) rstX：將原本的句子經過 RST (Rhetorical Structure Theory Discourse Treebank) 規則分類 [5]

(6) pdtbX：將原本的句子經過 PDTB (Rhetorical Structure Theory Discourse Treebank) 規則分類 [6]

(7) TFIDFX：將原本的句子經過詞頻 tf(term frequency)和逆向檔案頻率 idf (inverse document frequency) 規則分類

效能衡量指標(4 組)

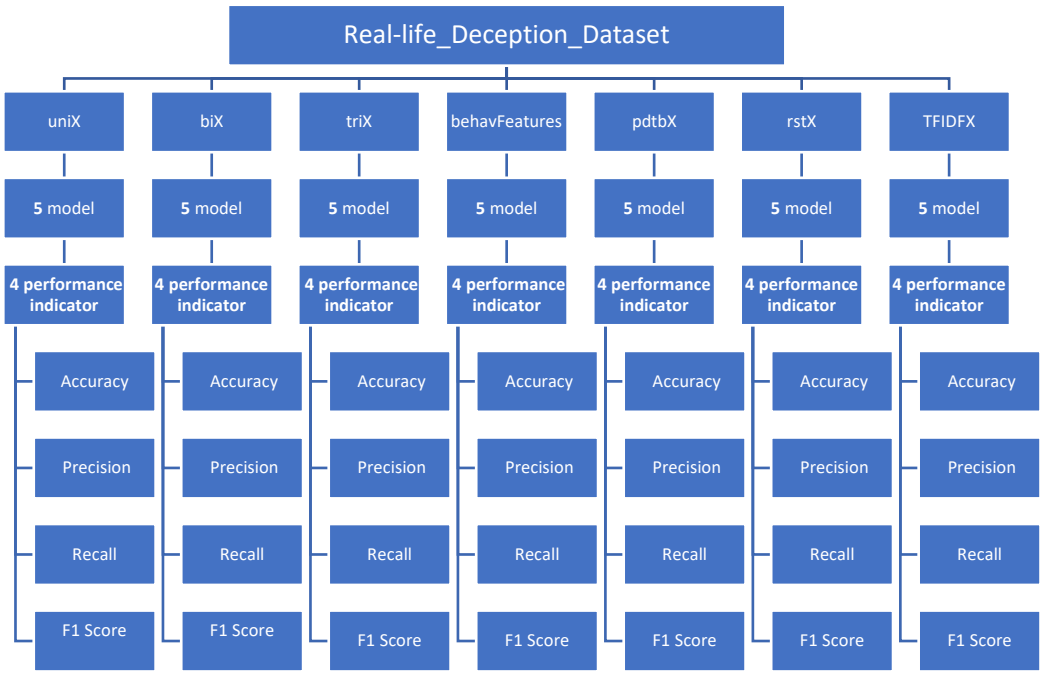
(1) 準確率(Accuracy)：總樣本中有幾個是預測正確的

(2) 精確率(Precision)：陽性的樣本中有幾個是預測正確的

(3) 召回率(Recall)：事實為真的樣本中有幾個是預測正確的

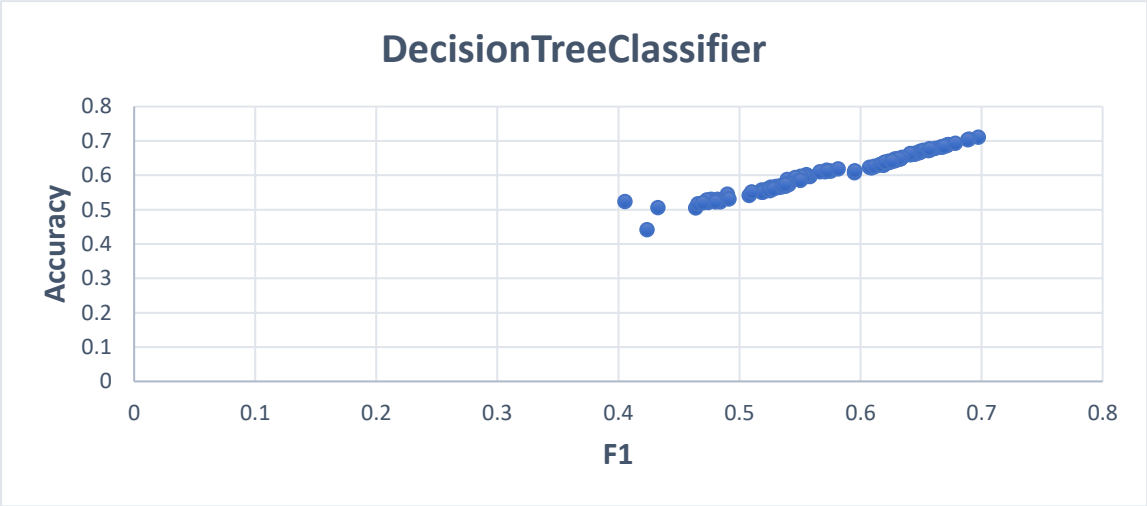
(4) F1 Score：精確率與召回率的調和平均數

實驗流程圖：

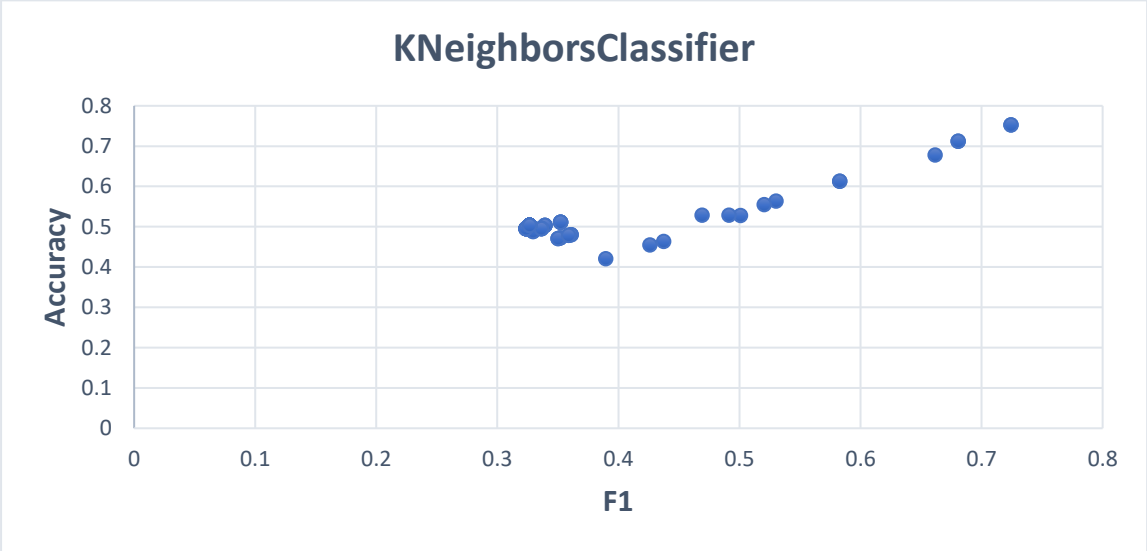


實驗結果(Accuracy):

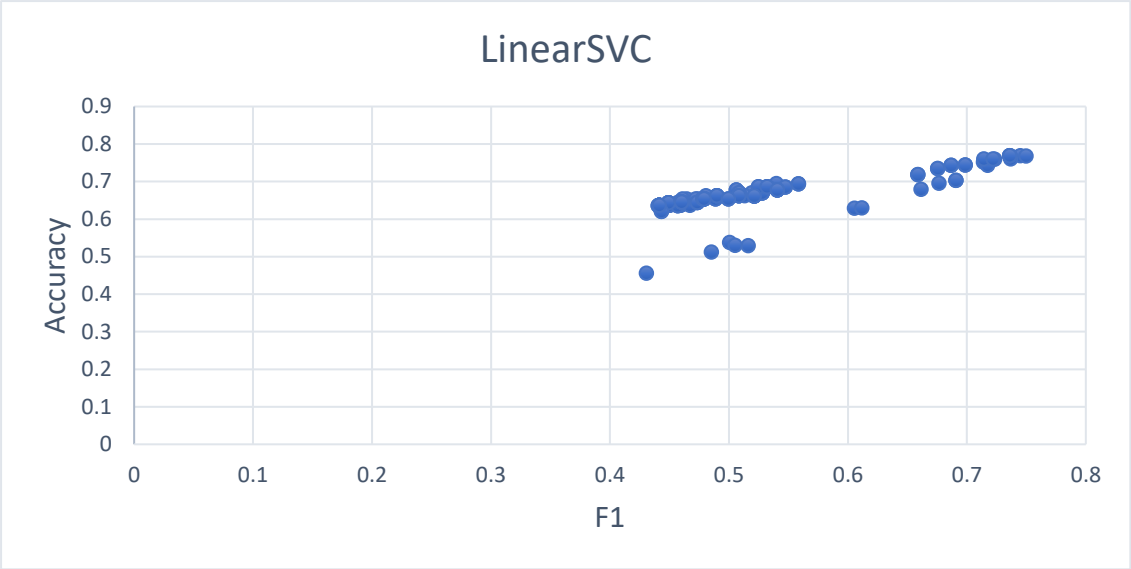
各圖片下方列出該 model 中最高的 Accuracy 發生的其他效能衡量指標



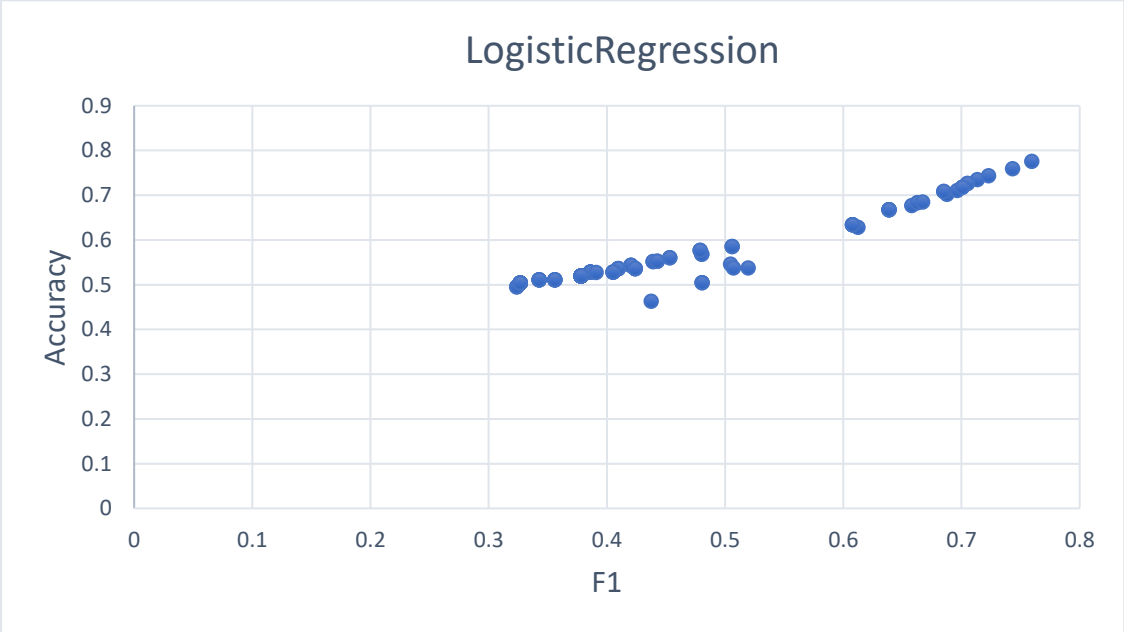
Feature Set	Accuracy	Precision	Recall	F1
triX_behavFeatures	0.7107051	0.7200079	0.7341904	0.697389868



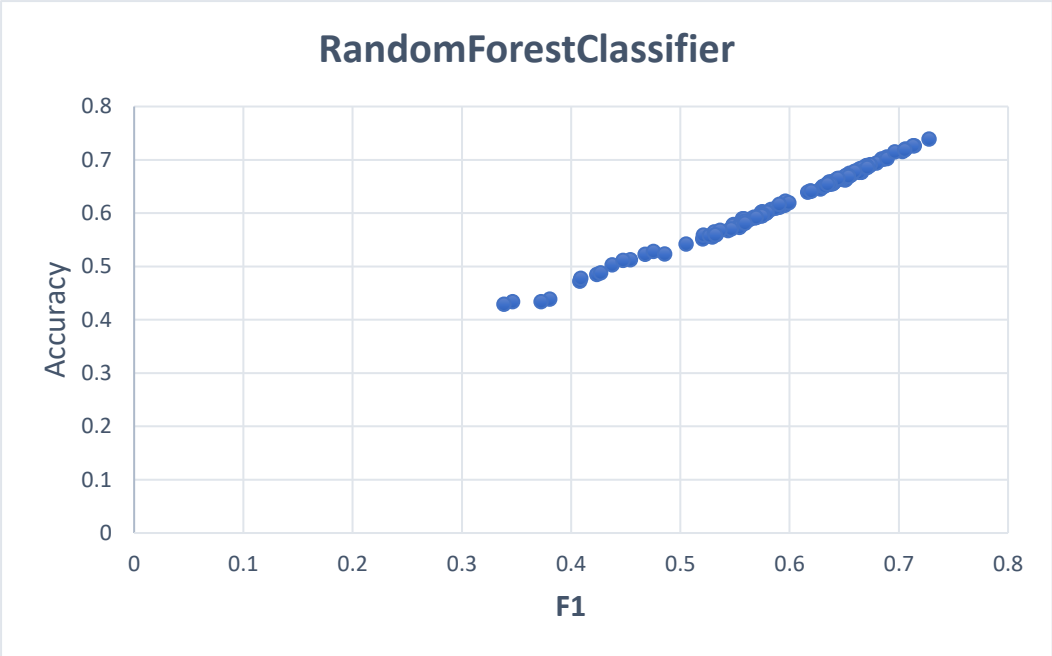
Feature Set	Accuracy	Precision	Recall	F1
behavFeatures	0.753205128	0.737640693	0.847301587	0.724302054



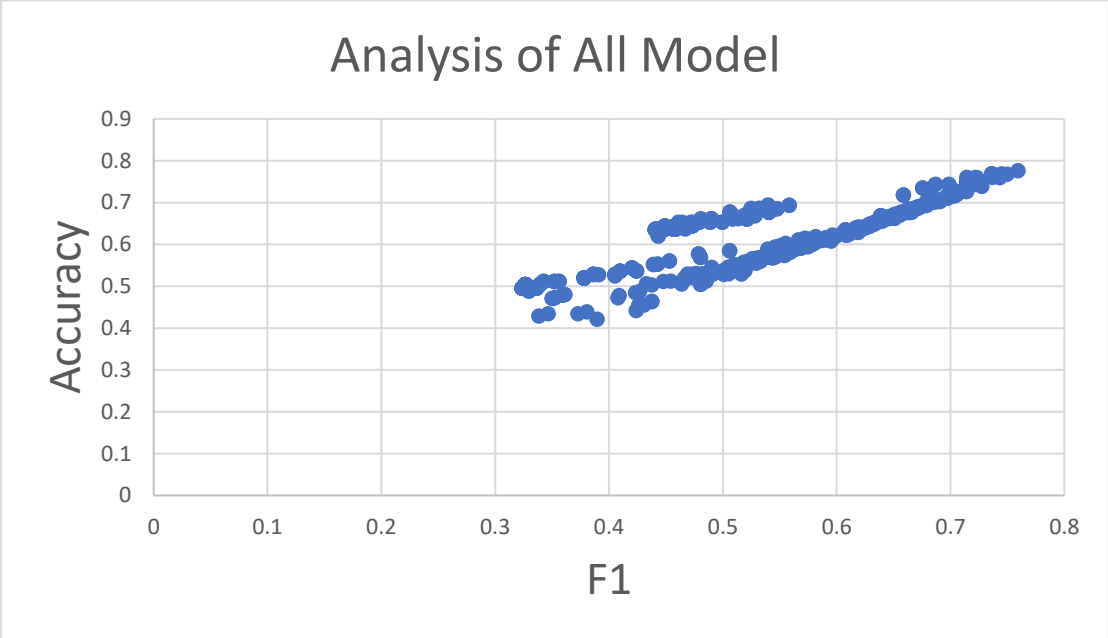
Feature Set	Accuracy	Precision	Recall	F1
behavFeatures	0.7685897	0.7506349	0.7941269	0.744980



Feature Set	Accuracy	Precision	Recall	F1
behavFeatures_pdtbX_TFIDFX	0.7762820	0.7396031	0.8250793	0.7593753



Feature Set	Accuracy	Precision	Recall	F1
behavFeatures_rstX	0.7390384	0.7482164	0.7840634	0.727617085



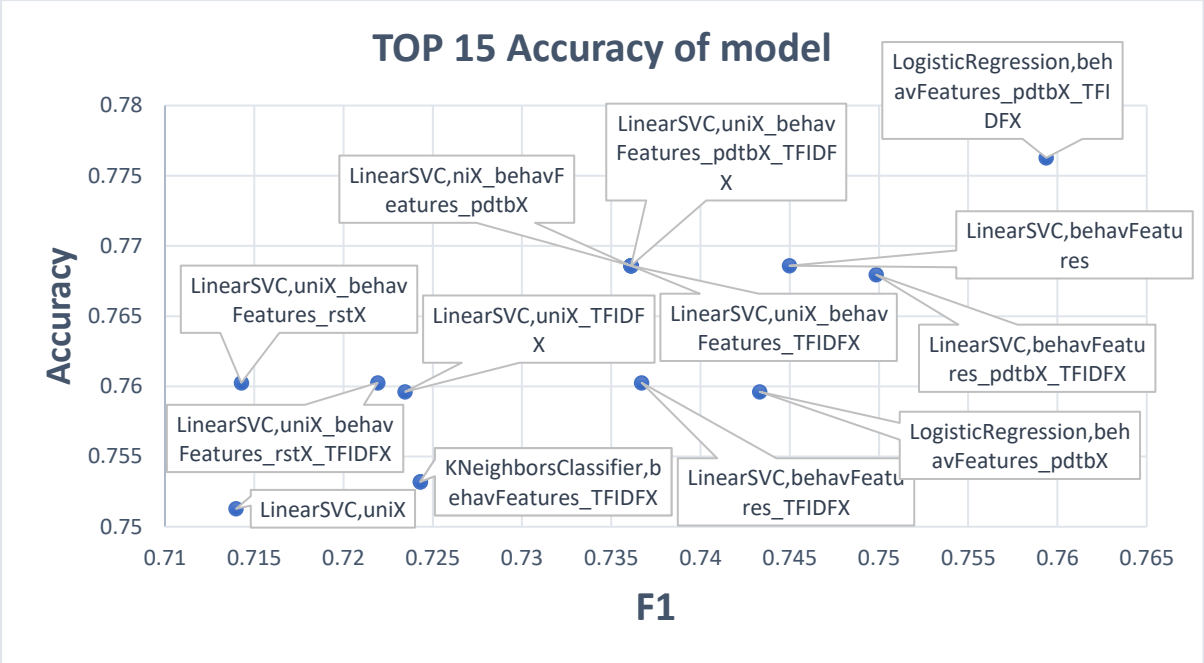
TOP 10 Accuracy of Model

Model	Feature Set	Accuracy	Precision	Recall	F1
LogisticRegression	behavFeatures_pdtbX_TFIDFX	0.776282	0.739603	0.825079	0.759375
LinearSVC	behavFeatures	0.76859	0.750635	0.794127	0.744981
LinearSVC	uniX_behavFeatures	0.76859	0.796151	0.682857	0.736109
LinearSVC	behavFeatures_pdtbX_TFIDFX	0.767949	0.736865	0.802857	0.74984
LinearSVC	behavFeatures_TFIDFX	0.760256	0.73873	0.794127	0.736706
LinearSVC	uniX_behavFeatures_rstX	0.760256	0.772143	0.66381	0.714276
LinearSVC	uniX_behavFeatures_rstX_TFIDFX	0.760256	0.786429	0.682857	0.721917
LinearSVC	uniX_TFIDFX	0.759615	0.814444	0.631905	0.723454
LogisticRegression	behavFeatures_pdtbX	0.759615	0.734643	0.797302	0.743314
KNeighborsClassifier	behavFeatures	0.753205	0.737641	0.847302	0.724302

Low 10 Accuracy of Model

Model	Feature Set	Accuracy	Precision	Recall	F1
KNeighborsClassifier	pdtbX_TFIDFX	0.464103	0.488824	0.592222	0.437475
LogisticRegression	pdtbX	0.463462	0.484048	0.56127	0.437527
KNeighborsClassifier	rstX_TFIDFX	0.455128	0.435952	0.44	0.426065
LinearSVC	pdtbX	0.455128	0.481667	0.550159	0.430582
DecisionTreeClassifier	rstX_pdtbX	0.441859	0.443214	0.370302	0.423731
RandomForestClassifier	triX_pdtbX	0.43891	0.369996	0.371619	0.380444
RandomForestClassifier	triX	0.43391	0.315667	0.328	0.34652
RandomForestClassifier	triX_rstX_pdtbX	0.433782	0.387004	0.33081	0.37252
RandomForestClassifier	triX_rstX	0.42891	0.299864	0.268143	0.338451
KNeighborsClassifier()	rstX	0.421154	0.473214	0.419524	0.389538

結果分析:



從結果可以清楚地看出，僅使用 behavFeatures,pdtbX 與 TFIDFX 的組合特徵集與原先數據經過 Logistic Regression 處理過經過 Kfold 交叉比對後的效果最佳，可以達到高達約 77.6% (0.776)的準確率，而使用 behavFeatures 與原先數據經過 LinearSVC 處理過經過 Kfold 交叉比對後的結果也得到 76.8%的準確率。

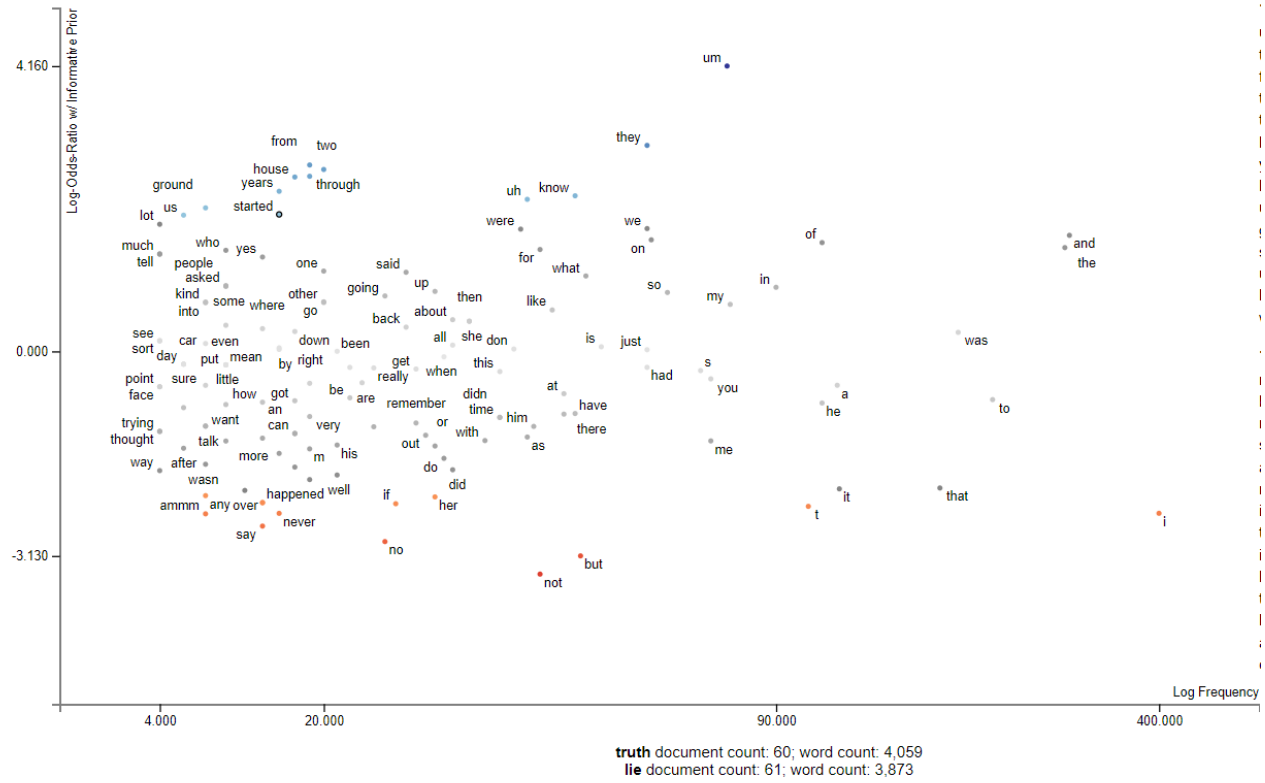
在前 10 名的 Accuracy of Model 中很明顯 behavFeatures 扮演一個很重要的角色，基本上每組 model 的最高準確率都會有使用 bigram 和 behavFeatures 作交叉比對,這也代表著 bigram 和 behavFeatures 包含了最大程度的文句含意,因而不會使的文字數據經過處理後失去太多的分析意義，而也從最低 Accuracy model 中可以發現經過 trigram 處理過後的文字數據可能會使的原本的語句意義消失或是變更，導致準確率只有相當低的 20.6%。

三、結論

在此專題的數據集分析中與原作者所的到的結果有許多相同，唯一不同的是我將數據集的處理擴大，並建立了一個有 2000 多筆實驗數據的資料庫，以提供評估 model 的完整性，也因為有著更多的額外參考數據，也可以評估中間數據的分布狀態是否合理。

在 rst 與 pdtb 的特徵中的許多特徵擁有相同的變化，並且在特徵分佈中的分佈基本相同，所以在經過 model 後有需多的準確率是相當接近的，而最後也建立了一個數據相當完善的資料庫，有著相當多的變化性。

最後我們觀察圖一後可以發現此資料庫內的大多單詞都集中於分析圖左半邊，這也使得再分析數據的部分有較少的參考資料，因此希望在未來可以收集更高質量的特徵數據集，或者需要收集更多的數據來更加有效率地提高在欺騙數據上的話語解析功效。



圖一、Real-Life Trial Data 的 dataframe 分析圖

（ X 軸為在 truth 時單詞所出現的數目，y 軸為在 lie 時單詞所出現的數目 ）

四、參考文獻

[1]W. G. Iacono, “Accuracy of polygraph techniques: Problems using confessions to determine ground truth,” Physiology & Behavior, vol. 95, no. 1, pp. 24–26, 2008, issn: 0031-9384. doi: 10.1016/j.physbeh.2008.06.001.

[2] V. P´erez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, “Deception Detection using Reallife Trial Data,” in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ser. ICMI ’15, New York, NY, USA: Association for Computing Machinery,Nov.2015,pp.59–66,isbn:978-1-4503-3912-4

[3] Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., & Webber, B. L. (2008, May). The Penn Discourse TreeBank 2.0. In LREC.

[4] Lynn Carlson, Daniel Marcu, and Mary EllenOkurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. Springer.

[5] ——, “A Linear-Time Bottom-Up Discourse Parser with Constraints and Post-Editing,” in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: LONG PAPERS), Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 511–521. doi: 10.3115/v1/P14-1048. [6] N. Okazaki, Crfsuite: A fast implementation of conditional random fields (crfs), 2007.

[6] YAN Weirong, XU Yang, ZHU Shanshan, HONG Yu, YAO Jianmin, ZHU Qiaoming. A Survey to Discourse Relation Analyzing. , 2016, 30(4): 1-11.

[7]Sen, Umut Mehmet, et al. "Multimodal Deception Detection using Real-Life Trial Data." IEEE Transactions on Affective Computing (2020).

[8] N. Okazaki, Crfsuite: A fast implementation of conditional random fields (crfs), 2007.

五、計畫管理與團隊合作方式

專題主要過程為實驗室內專題學長帶領學習，在第一學期的部分由於對於機械學習model的建構部分不太熟悉，因而花了一整個學期在學習如何處理資料集與學習模型的建構，除了學長給的教學資料外，也在學期內完成了一些零碎的與專題相關的資料處理作業，主要是熟悉如何將得到的資料庫作完善的處理與應用，並且在剩餘的時間內涉略些專題相關的論文，而學習的主要重點是背景分析和更多的測謊相關論文調查，在閱讀的論文中，也找到了一些學習的指導方針並也會與專題學長做定期的開會。

而第二學期的部分主要來到實作的部分，開始從網路上抓取建構好且完整數據集，並試著用他人建構好model試試看數據處理的部分，並也從中學習如何將學習到機械學習部分加以整合與應用在自己設計的model，在其中我覺得在文字數據處理的部分相對較難，畢竟在機械學習當中主要是以數字作為分析元件，而要如何將文字經過特定處理後再轉換為數字並丟入model內作分析是一件相當困難但是有趣的任務，在其中處理數據的部分由於原始數據量較小，因而才去想到去做一個特徵組合集以提高原始數據的變化性，要如何在設計特徵集中也要考慮如何不要做太大的變化，使原始數據保有其完整性。

而雖然許多數據集內通常會有許多已經標記好的特徵，但是要如何做出與他人不同的地方就是實做專題學習的意義，我也很謝謝學長提供了我大量的資料與安排學習規劃，讓我一步一步地慢慢開始接觸測謊這項大的學習領域，並且在我實作的報告或程式內提供我相對應的建議與教學。