類神經期末報告

SENTIMENT ANALYSIS USING TRANSFORMER BASED HYBRID MODELS

第五組: 李睿安 李文濠 郭恒綸 林偉盟

2026/6/12

CONTENTS

CONTENTS

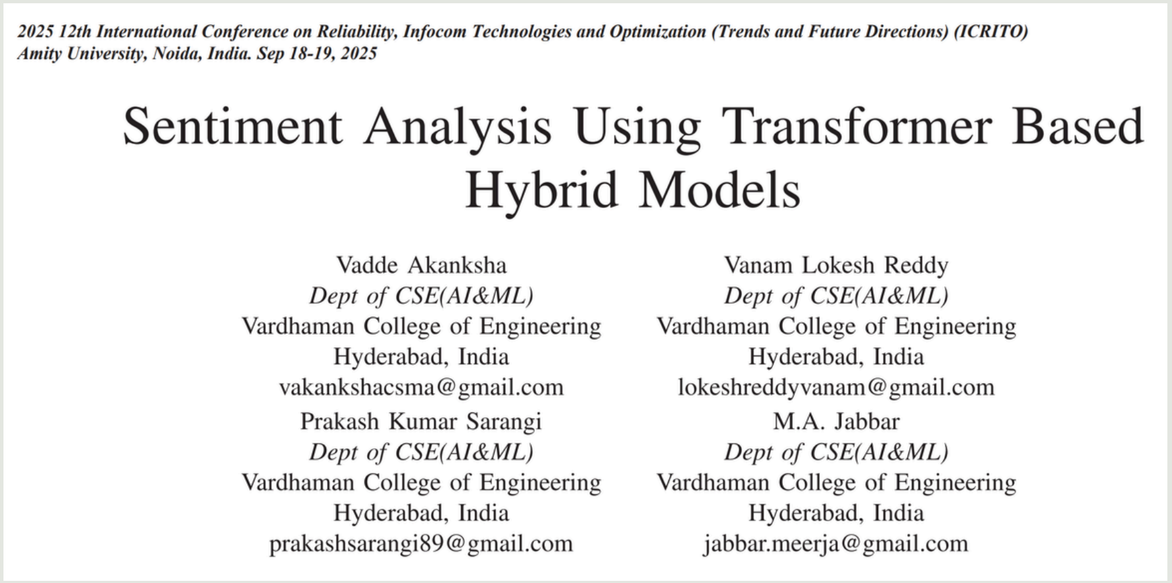
- **Literature Review**

1. **The main topic of this paper**
2. **Motivation, problem definition, and AS-IS/TO-BE analysis**
3. **Methods proposed to solve the problem**
4. **The main contribution of this work**

- **Group Topic**

1. **AS-IS**
2. **TO-BE**

THE MAIN TOPIC OF THIS PAPER 1/2



2025 12th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. Sep 18-19, 2025

Sentiment Analysis Using Transformer Based Hybrid Models

Vadde Akanksha
Dept of CSE(AI&ML)
Vardhaman College of Engineering
Hyderabad, India
vakankshacsma@gmail.com

Vanam Lokesh Reddy
Dept of CSE(AI&ML)
Vardhaman College of Engineering
Hyderabad, India
lokeshreddyvanam@gmail.com

Prakash Kumar Sarangi
Dept of CSE(AI&ML)
Vardhaman College of Engineering
Hyderabad, India
prakashsarangi89@gmail.com

M.A. Jabbar
Dept of CSE(AI&ML)
Vardhaman College of Engineering
Hyderabad, India
jabbar.meerja@gmail.com

這篇論文主要探討如何利用兩種 Transformer 模型：

RoBERTa 和 T5，來提升社群媒體文字的情緒分類效果。

任務是將 Twitter 文字分類為三種情緒類別：

1. Positive 正向
2. Neutral 中立
3. Negative 負向

論文使用 Kaggle 上的 Twitter Entity Sentiment Analysis Dataset，原始資料約有 74,994 筆英文推文，原本包含 positive、negative、neutral、irrelevant 四類，研究中將 irrelevant 類別移除，只保留三種主要情緒類別進行分類。資料切分方式為 70% training、10% validation、20% testing

THE MAIN TOPIC OF THIS PAPER 2/2

- 傳統情緒分析模型通常是單一模型，例如只使用 BERT、RoBERTa、T5，或是使用簡單的 ensemble，例如 majority voting 或 probability averaging
- 這些方法在處理模糊語句時，常常無法有效判斷。例如一則推文可能表面上是正向，但其實語氣偏中立，或是推文帶有抱怨但不明顯，容易在 neutral 和 negative 之間混淆

模型	角色	優點
RoBERTa	Discriminative classifier	擅長根據上下文做分類
T5	Generative sentiment model	擅長用 text-to-text 的方式產生情緒標籤，也具備較好的語意理解與可解釋性
Hybrid Model	Decision-level fusion	結合兩者輸出，降低單一模型判斷錯誤

MOTIVATION, PROBLEM DEFINITION, AND AS-IS/TO-BE ANALYSIS 1/3

- 近年來，社群媒體、電商評論、客服訊息、金融市場評論等文字資料大量增加
- 企業或研究者希望能自動分析這些文字中的情緒，了解使用者對產品、品牌、服務或事件的態度
- 論文指出，sentiment analysis 可應用於 e-commerce、customer service、social media websites 和 financial sectors，用來分析顧客意見、品牌聲譽與市場趨勢

困難	說明
Noisy and informal text	社群文字常包含網址、使用者標記、emoji 與非正式語氣
Contextual nuances	同一句話可能因上下文不同而呈現不同情緒
Ambiguous tweets	模糊或邊界語句容易造成情緒分類錯誤

困難	說明
Borderline sentiment	neutral / positive 或 neutral / negative 之間容易混淆
Single-model limitation	單一模型可能無法完整處理模糊語意與邊界情緒
Model disagreement	不同模型預測不一致時，傳統 majority voting 不一定能明確解決衝突

MOTIVATION, PROBLEM DEFINITION, AND AS-IS/TO-BE ANALYSIS 2/3

- 給定一段 Twitter 文字，模型需要自動判斷該文字的情緒類別是 positive、neutral 或 negative，並且希望在模糊、非正式、語意不清楚的推文中，也能降低分類錯誤　目標是提升：
  - Accuracy
  - Precision
  - Recall
  - F1-score
  - ROC-AUC
  - Confusion matrix 中各類別的穩定分類效果

MOTIVATION, PROBLEM DEFINITION, AND AS-IS/TO-BE ANALYSIS 3/3

分析項目	目前狀況 / 現有方法(AS-IS)	問題	論文提出的改善方式(TO-BE)
模型使用方式	多數研究使用單一 模型，例如 BERT、RoBERTa、T5	單一模型容易受到自身限制影響	結合 RoBERTa 和 T5，形成 hybrid model
Ensemble 方法	常見方法是 majority voting 或 probability averaging	當模型意見不一致時，無法明確解決衝突	使用 decision-level fusion 和 precedence rules 處理模型 disagreement
社群文字處理	Twitter 文字短、雜訊多、語氣模糊	neutral / positive、neutral / negative 容易分類錯	利用 T5 的語意生成能力與 RoBERTa 的分類能力互補
可解釋性	傳統分類模型通常只輸出 label	使用者難理解模型為什麼這樣判斷	T5 採 text-to-text 架構，未來可延伸成產生自然語言解釋
效能	單一模型雖然效果不錯，但仍有邊界錯誤	對模糊推文判斷不穩定	Hybrid model 在 accuracy、precision、recall、F1-score 上都高於單一模型

METHODS PROPOSED TO SOLVE THE PROBLEM 1/7

- **整體方法**
  - 準備 Twitter sentiment dataset
  - 進行文字前處理
  - 分別 fine-tune RoBERTa 和 T5
  - 將兩個模型的輸出進行 hybrid fusion
  - 使用 accuracy、precision、recall、F1-score、ROC-AUC、confusion matrix 評估模型效果

METHODS PROPOSED TO SOLVE THE PROBLEM 2/7

- **Dataset 資料集**

- 論文使用 Twitter Entity Sentiment Analysis Dataset，來源為 Kaggle。原始資料包含約 74,994 筆英文推文，類別包含：
  - Positive、Negative、Neutral、Irrelevant
  - 研究中將 irrelevant 類別移除，因為 irrelevant 不屬於主要情緒分類任務

Positive	Viking Assassin's Creed looks great and awesome.	維京題材的刺客教條看起來很棒、很精彩。
Negative	This is your fault guards for standing in my way.	這都是你們守衛的錯，因為你們擋在我的路上。
Neutral	Rock-Hard La Varlope, RARE & POWERFUL, HANDSOME JACKPOT, Borderlands 3.	這句主要是在描述遊戲物品或遊戲內容，沒有明顯正面或負面情緒。
Irrelevant	Clown check.	這句和 Assassin's Creed 主題沒有明顯關係，所以屬於不相關資料。

METHODS PROPOSED TO SOLVE THE PROBLEM 3/7

- **Data Preprocessing 文字前處理**
  - Lowercase :統一大小寫，降低文字差異
  - Unicode normalization :統一文字與 emoji 編碼
  - URL masking :將網址轉成 placeholder，避免網址造成雜訊
  - User tag masking :將 @user 轉成 placeholder，避免特定帳號干擾模型
  - Tokenization :將文字轉成模型可讀取的 token
  - Max length : = 128 限制輸入長度，超過則截斷
  - Remove irrelevant class :移除非情緒類別
  - Class-aware sampling :處理類別不平衡問題

METHODS PROPOSED TO SOLVE THE PROBLEM 4/7

- RoBERTa Model(Robustly Optimized BERT Pretraining Approach)

  - 移除 BERT 的 Next Sentence Prediction task
  - 使用更多資料訓練
  - 使用更大的 batch size
  - 使用 dynamic masking
  - 屬於 encoder-only Transformer
  - 適合做分類任務

RoBERTa 的訓練設定：

Component	Configuration
Model	RoBERTa-base
Encoder Layers	12
Classification Head	768 → 384 → 3
Activation Function	GELU
Loss Function	Categorical Cross-Entropy
Optimizer	AdamW
Learning Rate	5×10^{-5}
Epochs	3
Batch Size	32
Dropout Rate	0.1

METHODS PROPOSED TO SOLVE THE PROBLEM 5/7

- **T5 Model (Text-to-Text Transfer Transformer)**

- T5 的特色是把所有 NLP 任務都轉成 text-to-text task。也就是說，不管是翻譯、摘要、問答或分類，輸入和輸出都被視為文字
- T5 的優點是它不只是分類器，也具備生成能力，因此未來可以延伸成產生解釋，例如說明「為什麼這則推文被判斷為 negative」

T5 的訓練設定：

Component	Configuration
Model	T5-small
Input Format	"analyze sentiment: <tweet>"
Output	positive / neutral / negative
Loss Function	Negative log-likelihood loss
Optimizer	AdamW
Learning Rate	5×10^{-5}
Epochs	10
Batch Size	16
Gradient Accumulation	64
Label Smoothing	0.1

METHODS PROPOSED TO SOLVE THE PROBLEM 6/7

- Hybrid RoBERTa–T5 model

  - 模型使用 weighted soft-voting。T5 和 RoBERTa 都會輸出情緒類別的機率向量，接著用加權平均結合。最佳權重是透過 validation set 的 grid search 找到：T5 :0.6、RoBERTa :0.4
  - 當兩個模型出現明顯 disagreement，且情緒類別機率差異超過 0.25 時，論文會優先採用 T5 的結果，因為 T5 具備較強的生成式語意理解與 contextual explanation 能力。論文也提到此方法屬於 decision-level fusion，並透過 precedence rules 解決模型之間的衝突

METHODS PROPOSED TO SOLVE THE PROBLEM 7/7

Model	Accuracy	Precision	Recall	F1-score
Fine-tuned T5	95.77%	95.84%	95.78%	95.80%
Fine-tuned RoBERTa	93.84%	94.22%	93.91%	93.89%
Hybrid RoBERTa–T5	96.01%	96.13%	96.00%	96.05%

HYBRID MODEL 在四個指標上都是最高。相較於單獨使用 T5，HYBRID MODEL 的 ACCURACY 從 95.77% 提升到 96.01%；相較於 ROBERTA，則從 93.84% 提升到 96.01%。這表示 HYBRID MODEL 確實有助於提升整體分類表現

THE MAIN CONTRIBUTION OF THIS WORK

- Contribution 1：提出 RoBERTa + T5 的 Hybrid Sentiment Analysis Model
- Contribution 2：使用 Decision-Level Fusion 解決模型衝突
- Contribution 3：改善模糊推文的情緒分類
- Contribution 4：兼具準確率與可解釋性潛力

CONTENTS

- Literature Review
  1. The main topic of this paper
  2. Motivation, problem definition, and AS-IS/TO-BE analysis
  3. Methods proposed to solve the problem
  4. The main contribution of this work

- **Group Topic**
  1. **AS-IS**
  2. TO-BE

AS-IS

- 從論文缺失出發
  - Table IV Hybrid 96.01% Accuracy ≠ Page 6 混淆矩陣手動計算 (96.24%)
  - Precedence Rules Conflict: Negative > Neutral > Positive ≠Figure 1 -

    If Positive & neutral -> Positive
  - 實作與宣稱不符

- 改善論文的實作結果
- 驗證論文在不同文本應用上的表現

  - 驗證在高度依賴語意推理、反諷等挑戰性文本上的泛化能力

初步實驗結果 - T5論文版本

	precision	recall	f1-score	support
negative	0.6954	0.8401	0.7610	4210
neutral	0.6854	0.4008	0.5058	3381
positive	0.6894	0.7858	0.7345	3777
accuracy			0.6914	11368
macro avg	0.6901	0.6756	0.6671	11368
weighted avg	0.6905	0.6914	0.6763	11368

sizes train=39786 val=5684 test=11368 params total=60.5M

初步實驗結果 - ROBERTA論文版本

	precision	recall	f1-score	support
negative	0.9169	0.9352	0.9259	4210
neutral	0.8967	0.8912	0.8939	3381
positive	0.9101	0.8949	0.9024	3777
accuracy			0.9087	11368
macro avg	0.9079	0.9071	0.9074	11368
weighted avg	0.9086	0.9087	0.9086	11368

sizes train=39786 val=5684 test=11368 params total=60.5M

初步實驗結果 - HYBRID論文版本

```
=== Hybrid (paper soft voting) on processed/kaggle_val.csv ===
model           acc   f1_macro     prec   recall      auc
t5           0.8765     0.8764   0.8774   0.8780   0.9686
roberta      0.9649     0.9650   0.9650   0.9652   0.9955
hybrid       0.8753     0.8751   0.8762   0.8769   0.9762

disagreement: 99/826 (11.99%); T5 overrides after >0.25 gap: 98
```

```
Hybrid classification report:
              precision    recall  f1-score   support

    negative     0.8759    0.9321    0.9031       265
     neutral     0.8976    0.8000    0.8460       285
    positive     0.8552    0.8986    0.8763       276

    accuracy                         0.8753       826
   macro avg     0.8762    0.8769    0.8751       826
weighted avg     0.8765    0.8753    0.8745       826
```

sizes train=39786 val=5684 test=11368 params total=60.5M

初步實驗結果 - T5論文版本

評估指標	你的 T5 初步實驗結果 (Image 5)	論文宣稱的 T5 數據 (Table I) PDF	兩者差距
Accuracy（準確度）	69.14%	95.77%	落後 26.63%
Precision（精確度）	69.01% (Macro)	95.84%	落後 26.83%
Recall（召回率）	67.56% (Macro)	95.78%	落後 28.22%
F1-Score（F1 分數）	66.71% (Macro)	95.80%	落後 29.09%

改善論文的實作結果

	論文版	改良版
RoBERTa model	roberta-base	roberta-base
RoBERTa head	768 -> 384 -> 3 GELU	相同
RoBERTa epochs	3	3
RoBERTa lr	5e-5	5e-5
RoBERTa 額外設定	weight_decay=0、 warmup=0、 fp16=False	weight_decay=0.01、 warmup=0.06、 fp16=True

	論文版	改良版
T5 model	t5-small	t5-small
T5 max input length	128	96
T5 epochs	10	6
T5 optimizer	AdamW	Adafactor
T5 lr	5e-5	1e-3
T5 label smoothing	0.1	0.0

改良實驗結果 - T5

	precision	recall	f1-score	support
negative	0.9574	0.8817	0.9180	4210
neutral	0.9312	0.8728	0.9011	3381
positive	0.8364	0.9571	0.8927	3777
accuracy			0.9041	11368
macro avg	0.9084	0.9039	0.9039	11368
weighted avg	0.9094	0.9041	0.9046	11368

改良實驗結果 - ROBERTA

	precision	recall	f1-score	support
negative	0.9164	0.9238	0.9200	4210
neutral	0.8978	0.9012	0.8995	3381
positive	0.9067	0.8954	0.9010	3777
accuracy			0.9076	11368
macro avg	0.9069	0.9068	0.9068	11368
weighted avg	0.9076	0.9076	0.9076	11368

改良實驗結果 - HYBRID

```
=== Hybrid (precedence rule) on processed/kaggle_val.csv ===
model           acc   f1_macro     prec   recall
t5           0.9649     0.9653   0.9667   0.9650
roberta      0.9722     0.9721   0.9722   0.9722
hybrid       0.9625     0.9628   0.9648   0.9627

disagreement: 36/826 (4.36%); neg<->pos conflicts (T5-priority): 15
```

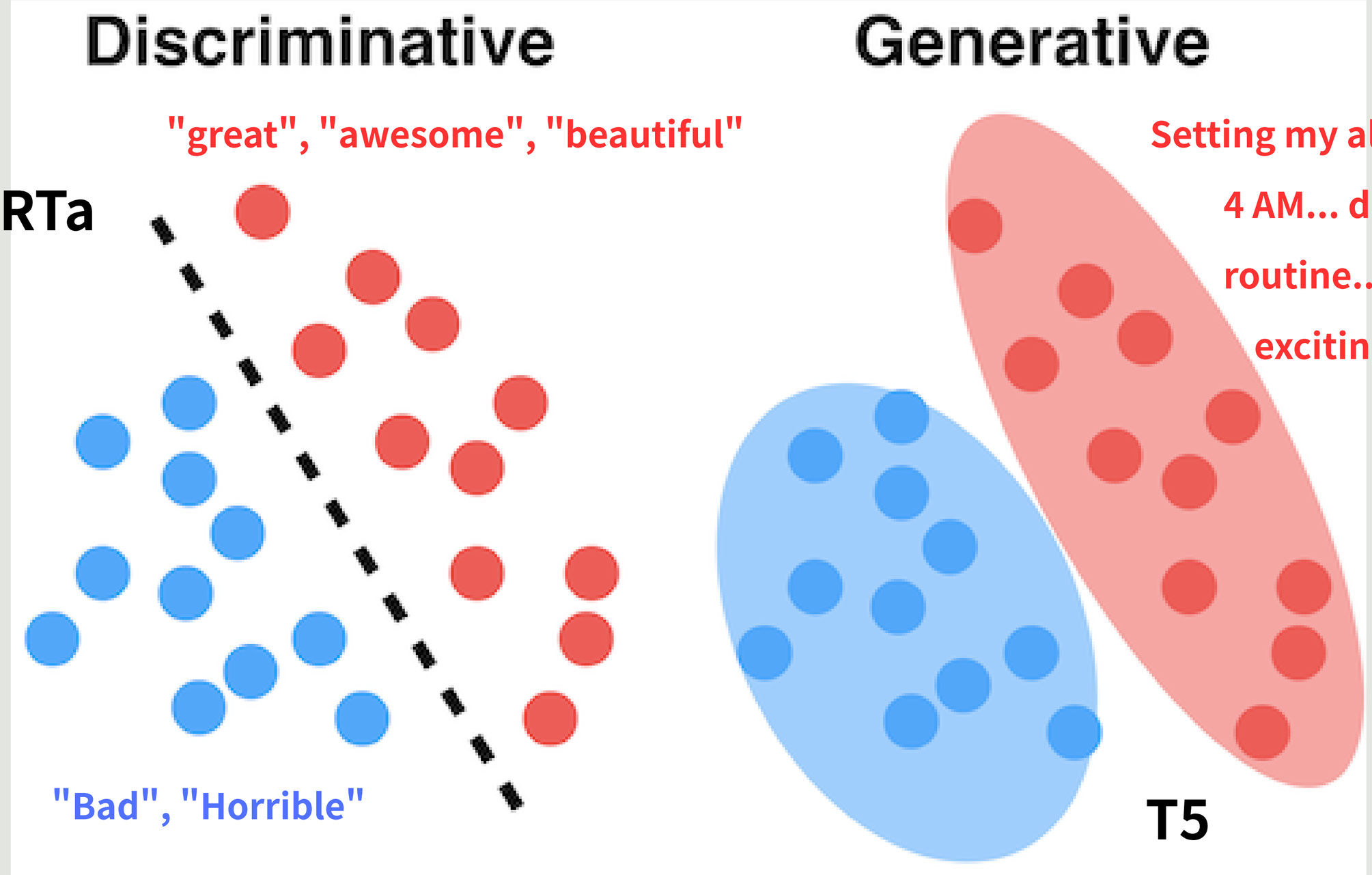
```
Hybrid classification report:
              precision    recall  f1-score   support

    negative     0.9883    0.9585    0.9732       265
     neutral     0.9925    0.9333    0.9620       285
    positive     0.9136    0.9964    0.9532       276

    accuracy                         0.9625       826
   macro avg     0.9648    0.9627    0.9628       826
weighted avg     0.9648    0.9625    0.9627       826
```

CONTENTS

- Literature Review
  1. The main topic of this paper
  2. Motivation, problem definition, and AS-IS/TO-BE analysis
  3. Methods proposed to solve the problem
  4. The main contribution of this work

- **Group Topic**
  1. AS-IS
  2. **TO-BE**

如何最大化凸顯Hybrid優勢



TweetEval irony

- 專門針對推特（Twitter）文本所設計的自然語言處理基準測試（Benchmark）

Dataset	筆數	二類別分布
Train	2862	irony 1445、non_irony 1417
Validation	955	non_irony 499、irony 456
Test	784	non_irony 473、irony 311

驗證論文在不同文本應用上的表現(範用性)

Irony(反諷):

- Would like to thank my nephew for giving me his horrible cold & sore throat etc.. Much appreciated!
- Double standards are always a fun thing.

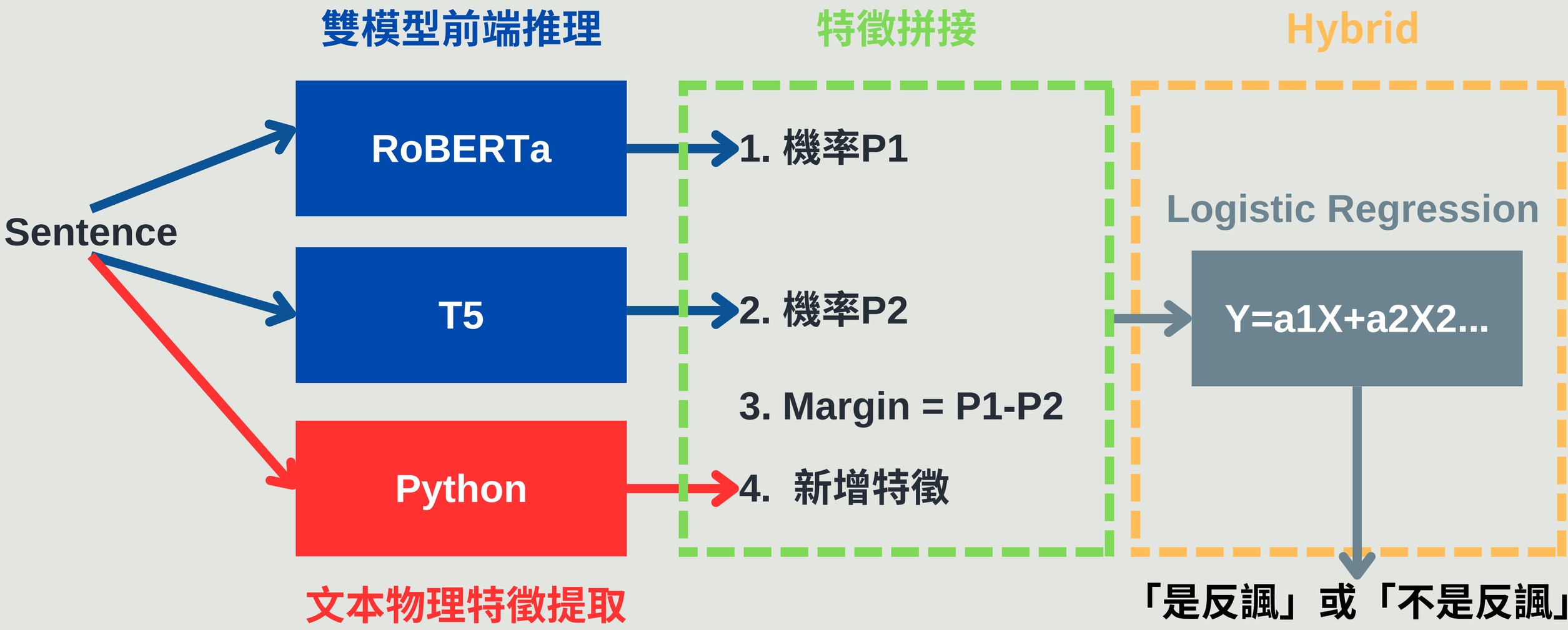
non_Irony(非反諷):

- On my lunch break so sleepy😴
- Simply having a wonderful christmas time :D

實驗設定

項目	改良版	Model For Irony
RoBERTa	roberta-base	twitter-roberta-base
RoBERTa head	768 -> 384 -> 3	768 -> 384 -> 2
T5	t5-small	google/flan-t5-small
T5 max target length	4	8
Hybrid	Rule-based precedence hybrid	Logistic Regression Combine
新增特徵	無	confidence、margin、文字長度、標點、hashtag、URL、mention、否定詞、笑聲詞、正負面 cue

Hybrid推論流程


雙模型前端推理
特徵拼接
Hybrid
Sentence
RoBERTa
T5
Python
1. 機率P1
2. 機率P2
3. Margin = P1-P2
4. 新增特徵
Logistic Regression
Y=a1X+a2X2...
文本物理特徵提取
「是反諷」或「不是反諷」

Hybrid Model Performance

Model	Accuracy	Macro F1	Macro Precision	Macro Recall	ROC-AUC
FLAN-T5-small	0.6939	0.6939	0.7270	0.7259	0.8076
Twitter-RoBERTa	0.7347	0.7324	0.7372	0.7476	0.8447
Rule Hybrid, T5 threshold 0.67	0.7372	0.7356	0.7428	0.7531	n/a
Stacked Hybrid + text features	0.7844	0.7784	0.7759	0.7839	0.8525

Conclusion

- 決策層的優先級融合（Precedence-rule Hybrid）
  - 為多模型融合提供了一個新的思維方向
- 本組貢獻與改良
  - 驗證論文的實作結果
  - 驗證論文在不同文本應用上的表現，並改良Hybrid

Division of Roles

- System Design: 李睿安 李文濠 郭恒綸 林偉盟
- Literature Review: 李睿安 李文濠 郭恒綸 林偉盟
- PPT Integration: 李睿安 李文濠 郭恒綸 林偉盟
- Presentation: 李睿安 李文濠 郭恒綸 林偉盟




Thank You