

110-2

# Natural Language Processing

Final Project

Yili Hsu, Chih Yao Chen

# Register to the competition !

1. 請至 [NLP期末專題分組](#) 確認小組名稱 (ex. Team\_0)
2. 到[NLP 110 Final Project - Basic \(Mandatory\)](#) 註冊並且點擊

**Join Competition**

# Register to the competition !

3. 進入 **Team** 即可編輯Team Name (請使用步驟一找到的小組名稱), 一組只需要執行一次即可

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Submit Predictions](#) [...](#)



Manage Team

Team Name

[Save Team Name](#)

This name will appear on your team's leaderboard position.

Team Members (1 of 5 maximum)


  **Yili Hsu** (you) Leader

# Register to the competition !

## 4. 使用Merge Teams將組員加入小組

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Submit Predictions](#) [...](#)

### Merge Teams

 Invite another team to merge with yours.

[Request Merge](#)

# Dataset - empathetic dialogues

Given different **situation** and **conversation**. Train your model to predict the **emotion (label)**.

**Label: Afraid**

**Situation:** Speaker felt this when...

“I’ve been hearing noises around the house at night”

**Conversation:**

**Speaker:** I’ve been hearing some strange noises around the house at night.

**Listener:** oh no! That’s scary! What do you think it is?

**Speaker:** I don’t know, that’s what’s making me anxious.

**Listener:** I’m sorry to hear that. I wish I could help you figure it out

**Label: Proud**

**Situation:** Speaker felt this when...

“I finally got that promotion at work! I have tried so hard for so long to get it!”

**Conversation:**

**Speaker:** I finally got promoted today at work!

**Listener:** Congrats! That’s great!







**Speaker:** Thank you! I’ve been trying to get it for a while now!

**Listener:** That is quite an accomplishment and you should be proud!

Paper ref: <https://arxiv.org/pdf/1811.00207.pdf>

# Dataset - empathetic dialogues

請至 [data](#) 下載 training set/ validation set/ testing set




< <b>train.csv</b> (16.07 MB)						
Detail Compact Column						
<b>About this file</b>						
This file does not have a description yet.						
∞ id	Δ conv_id	# utterance_idx	Δ prompt	Δ utterance	# label	
 0 76.7k	 <b>17844</b> unique values	 1 8	 <b>17578</b> unique values	 <b>75136</b> unique values	 0 31	
0	hit:0_conv:1	1	I remember going to the fireworks with my best friend. There was a lot of people_comma_ but it only ...	I remember going to see the fireworks with my best friend. It was the first time we ever spent time ...	13	
1	hit:0_conv:1	2	I remember going to the fireworks with my best friend. There was a lot of people_comma_ but it only ...	Was this a friend you were in love with_comma_ or just a best friend?	13	

# Dataset - empathetic dialogues

- id: 樣本的編號
- conv\_id: 一輪完整對話的編號 (提示: 可以對這個欄位 groupby)
- utterance\_idx: 在某一輪對話中, 表示第幾句話的編號
- prompt: 對於說話者情境的描述 (提示: hard baseline 以下可以只考慮此欄位)
- utterance: Speaker 或 Listener 在一輪對話中所講的話
- label: 整個對話的 emotion label

# Baselines

1. **Simple Baseline:** word frequency / tf-idf / Logistic regression
2. **Medium Baseline:** word2vec / GloVe / CNN / RNN-based model
3. **Hard Baseline:** BERT / RoBERTa / all turn conversation

#	Team	Members	Score	Entries	Last	Code
	hard.csv		0.59663			
	medium.csv		0.54765			
	simple.csv		0.45243			



# Evaluation Metric

- We use macro F1 score to evaluate the testing set performance.

$$F1 = 2 \frac{p \cdot r}{p + r} \text{ where } p = \frac{tp}{tp + fp}, \quad r = \frac{tp}{tp + fn}$$

- Macro F1 is the unweighted mean of F1 score for each class.
- Scikit-learn provides both implementation and documentation. ([ref.](#))

# Emotion Class to label mapping

- The following dictionary is used to map emotion class to its label.

```
{'sad': 0, 'trusting': 1, 'terrified': 2, 'caring': 3, 'disappointed': 4,  
'faithful': 5, 'joyful': 6, 'jealous': 7, 'disgusted': 8, 'surprised': 9,  
'ashamed': 10, 'afraid': 11, 'impressed': 12, 'sentimental': 13,  
'devastated': 14, 'excited': 15, 'anticipating': 16, 'annoyed': 17, 'anxious': 18,  
'furious': 19, 'content': 20, 'lonely': 21, 'angry': 22, 'confident': 23,  
'apprehensive': 24, 'guilty': 25, 'embarrassed': 26, 'grateful': 27,  
'hopeful': 28, 'proud': 29, 'prepared': 30, 'nostalgic': 31}
```



# 提醒事項

## NLP 110 Final Project - Basic (Mandatory)

- **請勿使用訓練集以外的資料來訓練模型**

包含：額外的文字、語料庫、知識圖譜等資源

- **請勿上網搜尋或「偷看」測試集的答案**

我們會根據繳交的程式碼來評估模型的表現是否合理

如有偷看答案或作弊等影響公平性的行為**一律以零分計算**

- **需繳交程式碼(包含參數的設定、random seed等)，請確保程式碼可以直接還原預測的結果，供助教執行並且檢查模型表現是否和程式碼吻合**

# 加分比賽(選擇性參加)

## NLP 110 Final Project - Advanced (Optional)

- **可以使用訓練集以外的資料來訓練模型**

包含:額外的文字、語料庫、知識圖譜等資源

- **請勿上網搜尋或「偷看」測試集的答案**

我們會根據繳交的程式碼來評估模型的表現是否合理

如有偷看答案或作弊等影響公平性的行為**一律以零分計算**

- **需繳交程式碼(包含參數的設定、random seed等), 請確保程式碼可以直接還原預測的結果, 供助教執行並且檢查模型表現是否和程式碼吻合**

# 書面報告繳交

- 請使用ACL template，請參考網址 (<http://acl2020.org/downloads/acl2020-templates.zip>) 的引導，可以使用 docx 檔或是 [overleaf](#) 進行撰寫。
- 頁數限制4到8頁(含)，使用英文撰寫。
- 依照 paper 撰寫格式依序撰寫：Introduction, Related Work, (Data), Method, Experiment, Conclusion, Work Division, Question and Answer, Reference。
- 不需寫 Abstract, 頁數限制不包含 Work Division, Question and Answer 與 Reference。
- Work Division請描述**小組內分工狀況**
- 請記得包含線上報告時同學和老師問題的回覆 (即 Qestion and Answer)。
- 將報告轉成 PDF 檔再繳交。

# Important deadlines

- Kaggle competition: **6/9 23:59pm (UTC+8)** 前可Submit
- 期末專題線上報告: **6/10, 6/17** 上課時間
- 程式碼繳交: **6/10 23:59pm**
  - 原則上只需要繳交一份, 若有參加加分比賽則繳交兩份程式碼。
- 書面報告繳交: **6/24 23:59pm**