Text Sentiment Classification with Prompt Learning

B092040016 陳昱逢

Assignment 6

1 Task 1

任務一分別測試了 BERT 和 RoBERTa 兩個模型。此任務首先採取預訓練好的模型直接 fine-tune 在 sentiment classification 的 task 上，我最後採用的模型為 RoBERTa-large 並在進入最後一層前加了 Dropout 來防止模型過擬合。Table 1 顯示了訓練 Epoch 為 5 的測試集結果。

Table 1: Simulation results of sentiment classification on Twitter US Airline

Model	**Accuracy**
BERT	0.8392
RoBERTa-base	0.8532
RoBERTa-large + Dropout	**0.8652**

為了再提升準確度，我訓練更久，Epoch 為 15，隨後利用 ensemble 的概念，選取 5 個 model 做權重投票投出結果，發現可以再有些微的提升。Table 2 呈現測試集結果。

Table 2: Simulation results of sentiment classification on Twitter US Airline

Model	**Accuracy**
RoBERTa-large + Dropout	0.8717
RoBERTa-large + Dropout + ensemble	**0.8722**

2 Task 2

任務二嘗試了三種不同的 template 並實驗於 zero-shot, one-shot and few-shot 三種情況中。Table 3 呈現我此次實驗不同 template 的設定。

```
self.one_shot = [
'I gave you one more try. Figured you could get a 1 hr flight right. Nope. Delayed an hr. Seems to be every time. It is negative.',
'I gave you one more try. Figured you could get a 1 hr flight right. Nope. Delayed an hr. Seems to be every time. The sentiment of this sentence is negative.',
'I gave you one more try. Figured you could get a 1 hr flight right. Nope. Delayed an hr. Seems to be every time. How do you feel after reading this sentence? I feel negative.'
                ]
```

Figure 1: one-shot prompts for different templates

Table 3: Manuallly designed template

Template	Prompt
Template 1	It is **[MASK]**.
Template 2	The sentiment of this sentence is **[MASK]**.
Template 3	How do you feel after reading this sentence? I feel **[MASK]**.

2.1 zero-shot

Table 4 呈現不同 template 在 zero-shot 情況下的實驗結果。

Table 4: Simulation results of zero-shot with different templates

Template	Accuracy	Precision	Recall	F1 score
Template 1	**0.4255**	**0.6375**	**0.4255**	**0.4236**
Template 2	0.2175	0.6013	0.2175	0.1504
Template 3	0.1616	0.4440	0.1616	0.0453

2.2 one-shot

Figure 1 展示我此次實驗針對不同 template 設定的 one-shot 範例提示。Table 5 呈現不同 template 在 one-shot 情況下的實驗結果。

Table 5: Simulation results of one-shot with different templates

Template	Accuracy	Precision	Recall	F1 score
Template 1	**0.3740**	**0.5388**	**0.3740**	**0.3667**
Template 2	0.1614	0.0260	0.1614	0.0449
Template 3	0.1614	0.0260	0.1614	0.0449

2.3 few-shot

Few-shot 則從訓練資料裡多拿幾個樣本來當做 prompt，我拿了三個樣本，主要想法是每一個類別都拿一個樣本。Table 6 呈現不同 template 在 few-shot 情況下的實驗結果。

3 Task 3

任務三實驗不同的手動 template 設定以及不同數量的 demonstration 的表現度。Table 7 呈現了三種不同的手動 template 的設定。

Table 6: Simulation results of few-shot with different templates

Template	Accuracy	Precision	Recall	F1 score
Template 1	**0.6270**	0.3931	**0.6270**	**0.4832**
Template 2	0.2631	0.5826	0.2631	0.2269
Template 3	0.3198	**0.6115**	0.3198	0.3038

Table 7: Manuallly crafted template

Template	Prompt
Manual Template 1	{" placeholder":"text_a"} It was {"mask"}.
Manual Template 2	{"placeholder":"text_a"} The sentiment of this sentence is {"mask"}.
Manual Template 3	{"placeholder":"text_a"} How do you feel after reading this sentence? I feel {"mask"}.

3.1 Different manually crafted templates

Table 8 比較了三種不同的手動 template 以及自動產生 template 的表現度，由結果可以觀察出自動生成 template 的表現效果達到最好。

Table 8: Performance comparison of different manually templates

Template	accuracy
Manual Template 1	0.6918
Manual Template 2	0.8623
Manual Template 3	0.7770
Auto-generate Template	**0.9082**

3.2 Differenet numbers of demonstrations

Table 9 比較了不同數量的 demonstrations 對模型表現度的影響，由結果可以觀察出使用 6 個數量的 demonstration 結果較好。以下每一段分別匯報對不同數量的 demonstration，最好的 template 跟 verbalizer。

對於 num_of_demonstration 為 1, best template 為 {"placeholder": "text_a"} It was {"mask"} . flat.nothing happens , and it happens to flat characters . It was terrible. a crisp psychological drama (and) a fascinating little thriller that would have been perfect for an old " twilight zone " episode . It was great.，而 best verbalizer 為 ['horrifying', 'terrifying']

對於 num_of_demonstration 為 6, best template 為 {"placeholder": "text_a"} It was {"mask"} . it was terrible.just a collection of this and that – whatever fills time – with no unified whole . It was terrible. serious movie-goers embarking upon this journey will find that the road to perdition leads to a satisfying destination . It was great. in that setting , their struggle is simply too ludicrous and borderline

Table 9: Performace comparison of different numbers of demonstrations

# of demonstraions	accuracy
1	0.9082
6	**0.9131**
8	0.9066

insulting . It was terrible. shyamalan takes a potentially trite and overused concept (aliens come to earth) and infuses it into a rustic , realistic , and altogether creepy tale of hidden invasion . It was great. nothing happens , and it happens to flat characters . It was terrible. serious movie-goers embarking upon this journey will find that the road to perdition leads to a satisfying destination . It was great. often lingers just as long on the irrelevant as on the engaging , which gradually turns what time is it there ? It was terrible. my big fat greek wedding is not only the best date movie of the year , it 's also a – dare i say it twice – delightfully charming – and totally american , i might add – slice of comedic bliss . It was great. the big finish is a bit like getting all excited about a chocolate eclair and then biting into it and finding the filling missing . It was terrible. a haunting tale of murder and mayhem . It was great. the plot is nothing but boilerplate clichés from start to finish , and the script assumes that not only would subtlety be lost on the target audience , but that it 's also too stupid to realize that they 've already seen this exact same movie a hundred times It was terrible. the filmmakers ' eye for detail and the high standards of performance convey a strong sense of the girls ' environment . It was great. ，而 best verbalizer 為 ['awful', 'good']

對於 num_of_demonstration 為 8, best template 為 {"placeholder": "text_a"} It was {"mask"} ..a rude black comedy about the catalytic effect a holy fool has upon those around him in the cutthroat world of children 's television . It was terrible. a crisp psychological drama (and) a fascinating little thriller that would have been perfect for an old " twilight zone " episode . It was great. as lively an account as seinfeld is deadpan . It was terrible. sweet and memorable film . It was great. just a collection of this and that – whatever fills time – with no unified whole . It was terrible. the film jolts the laughs from the audience – as if by cattle prod . It was great. just because a walk to remember is shrewd enough to activate girlish tear ducts does n't mean it 's good enough for our girls . It was terrible. (ramsay) visually transforms the dreary expanse of dead-end distaste the characters inhabit into a poem of art , music and metaphor . It was great. so we got ten little indians meets friday the 13th by way of clean and sober , filmed on the set of carpenter 's the thing and loaded with actors you 're most likely to find on the next inevitable incarnation of the love boat . It was terrible. diggs and lathan are among the chief

reasons brown sugar is such a sweet and sexy film . It was great. as lively an account as seinfeld is deadpan . It was terrible. a crisp psychological drama (and) a fascinating little thriller that would have been perfect for an old " twilight zone " episode . It was great. often lingers just as long on the irrelevant as on the engaging , which gradually turns what time is it there ? It was terrible. sweet and memorable film . It was great. but it would be better to wait for the video . It was terrible. (ramsay) visually transforms the dreary expanse of dead-end distaste the characters inhabit into a poem of art , music and metaphor . It was great. ，而 best verbalizer 為 ['well', 'tremendous']
Figure 2 呈現不同數量的 demonstrations 的準確度。

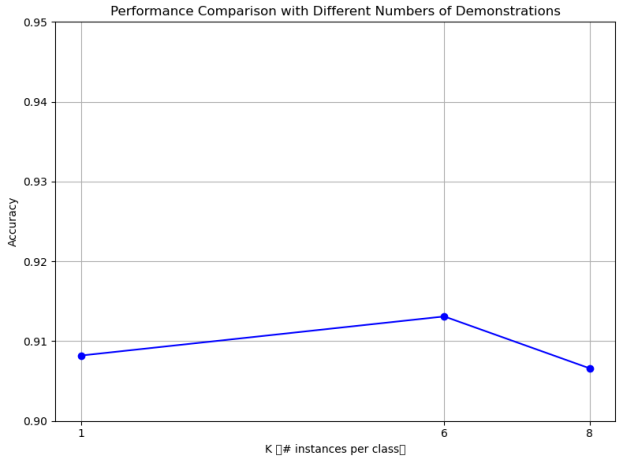

Figure 2: LM-BFF for different numbers of demonstrations (# instances per class)