Name: 陳以瑄 Student ID: 109705001

1. **Describe your understanding and findings about the attention mechanism by exBERT.**

   - Understanding
     attention mechanism 是指 encoder 在每個階段，會將輸入編成「不同的」特徵給 decoder。

   - Findings
     我選用的 model 為 distilbert-base-uncased
     我輸入的句子為“ The boy climbed trees, exploring the forest with his loyal dog.”

     Layer 1

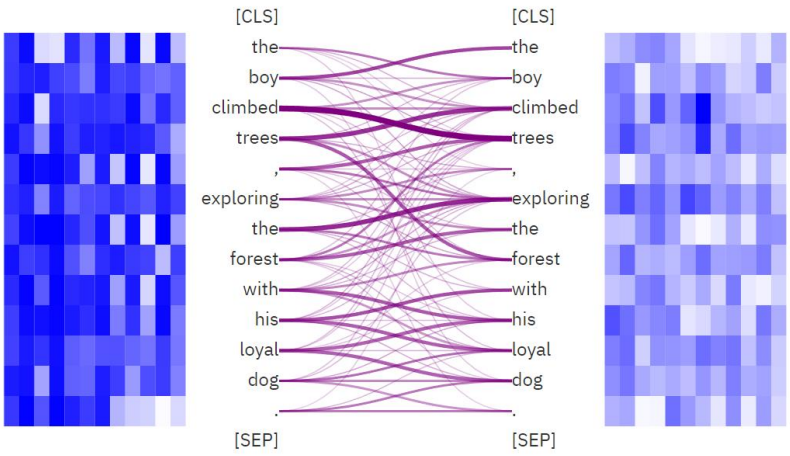


圖 1、Layer 1

可以發現右邊的字最主要都是受到前後兩個字的影響(線比較粗、顏色深)。以 climbed 為例，可以發現越接近的字越重要。

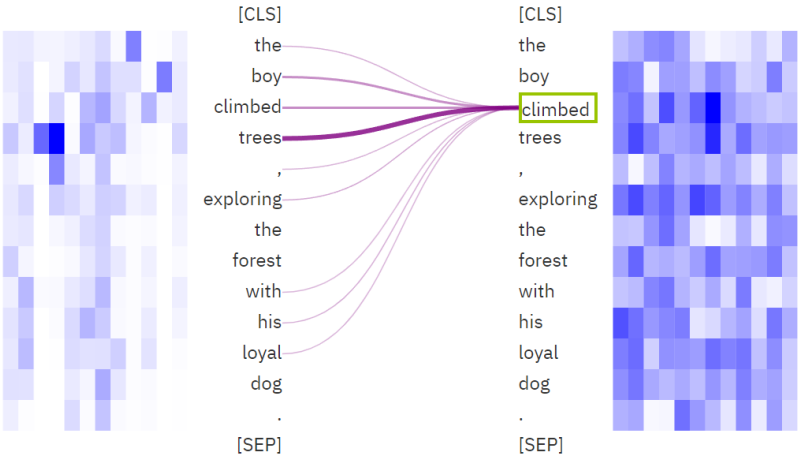


圖 2、Layer 1 “climbed”

而前後兩個字的影響力可以從 head 4 與 head 11 發現。

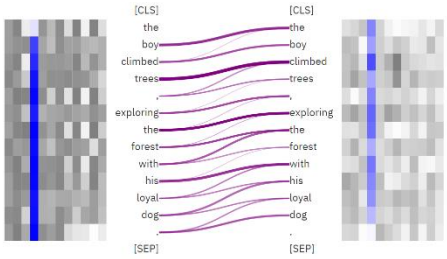


圖 3、Layer 1 head 4

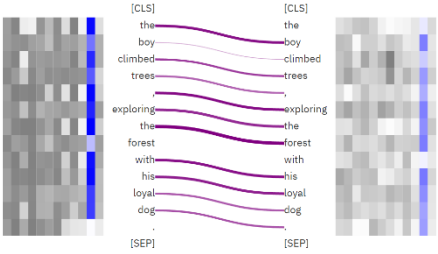


圖 4、Layer 1 head 11

Layer 2

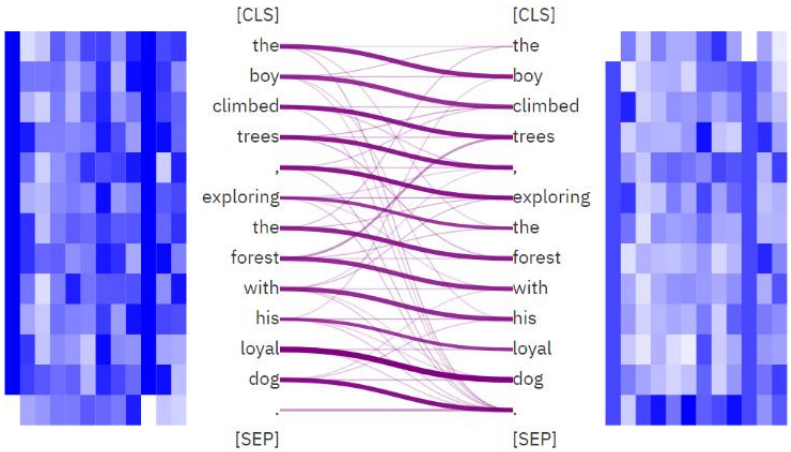


圖 5、Layer 2

可以發現右邊的字最主要都是受到<u>前一個字</u>的影響。這可以從 head 4 與 head 11 發現。

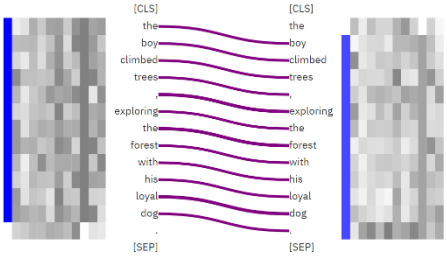


圖 6、Layer 2 head1

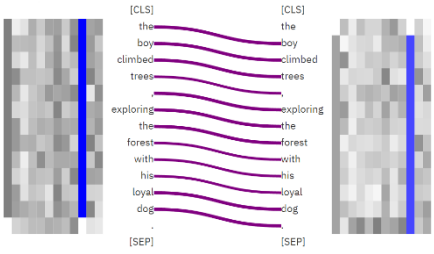


圖 7、Layer 2 head10

Layer 3

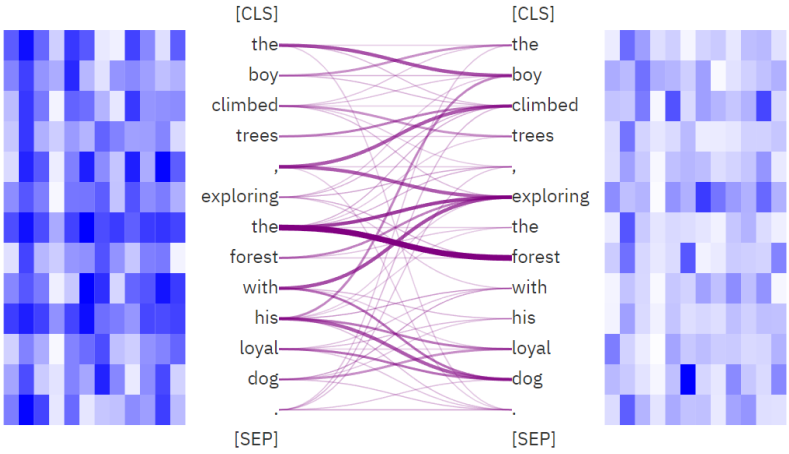


圖 8、Layer 3

從右邊比較深色的字：boy, climbed, exploring, forest, dog，我推測這層主要是針對名詞與動詞。

從圖 9 可以看到，在 head 4 左邊的 boy 與 his 這類與男生有關的詞彙影響著 boy。從圖 10 可以看到如果把 boy 換成 woman，並且把 his 換成 her，也會有類似的結果。因此我推測在 layer3 的 head 4 主要會抓出與主詞性別相關的 feature。

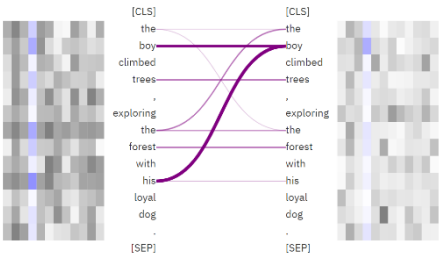


圖 9、Layer 3 head 4 “boy”

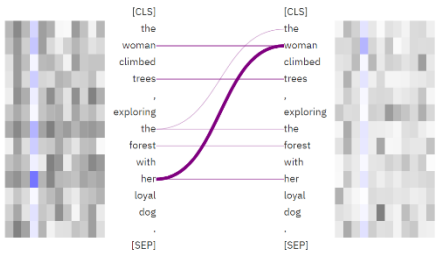


圖 10、Layer 3 head 4 “woman”

從圖 11 可以看到，在 head 6 右邊顏色最深的是像 forest 與 dog 的受詞。而從圖 12 會發現，在 head 7、8、11 右邊顏色最深的動詞，會受到受詞的影響。

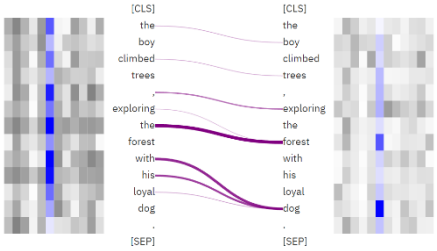


圖 11、Layer 3 head 6

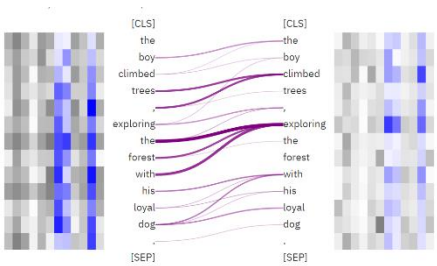


圖 12、Layer 3 head 7、8、11

Layer 4

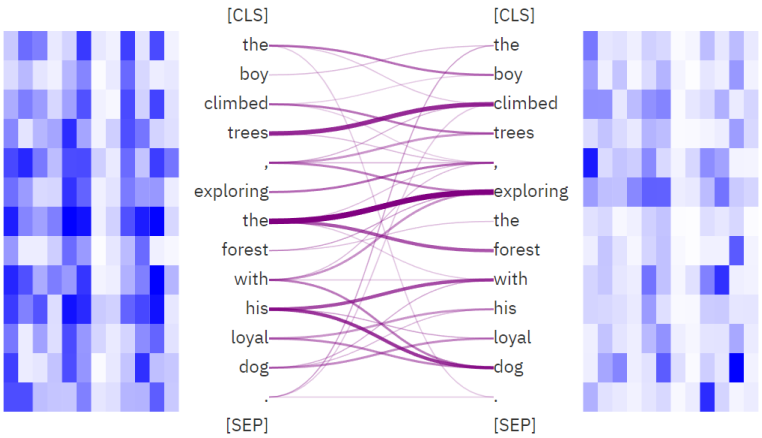


圖 13、Layer4

從右邊比較深色的字： climbed, exploring，我推測這層主要針對動詞。

從圖 14 可以看到，在 layer 4 head 10 其實與上層的結果類似，動詞會受到後續受詞的影響。

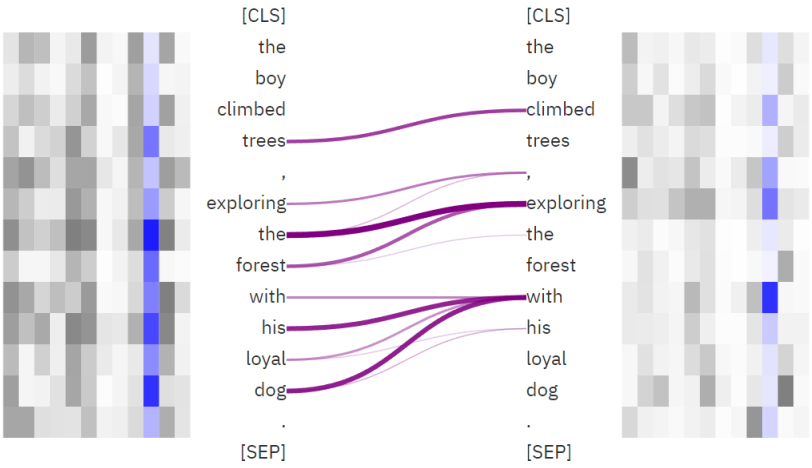


圖 14、Layer 4 head 10

Layer 5

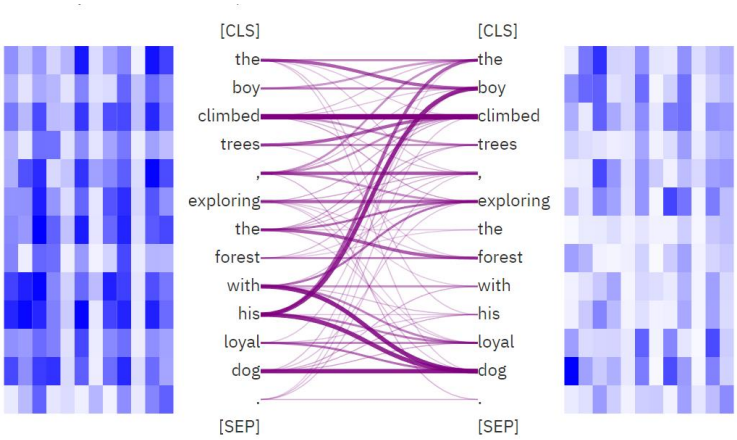


圖 15、Layer5

從上圖可以發現幾乎每個字都受到相同位置的字影響。這些水平線可以在 head7 與 head9 看到，不過 head7 主要針對標點符號。

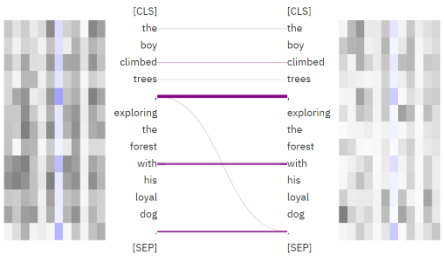


圖 16、Layer 5 head 7

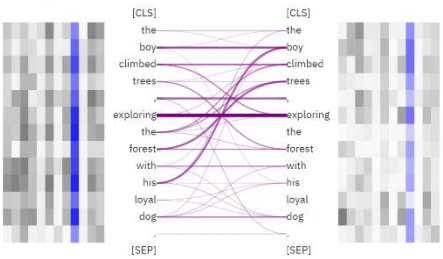


圖 17、Layer 5 head 9

Layer 6

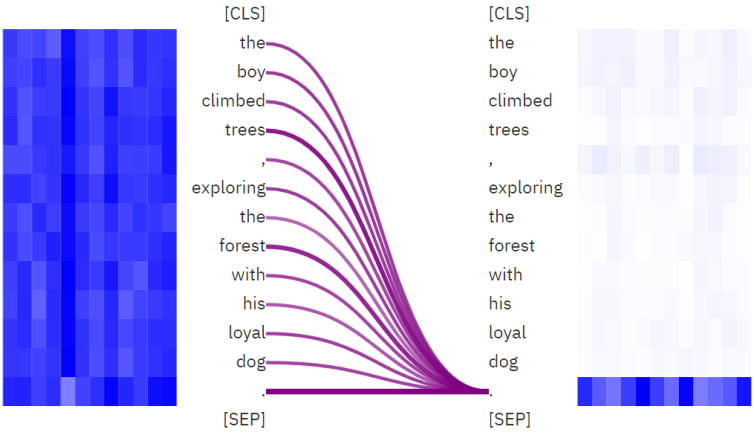


圖 18、Layer6

從上圖可以發現 layer6 最主要是針對句號。

2. **Compare at least 2 sentiment classification models**

我比較了助教提供的 TA_model_1.pt 與 TA_model_2.pt。

- 模型架構的比較

  TA_model_1 是採用 DistilBertModel，embedding 成 768 features；

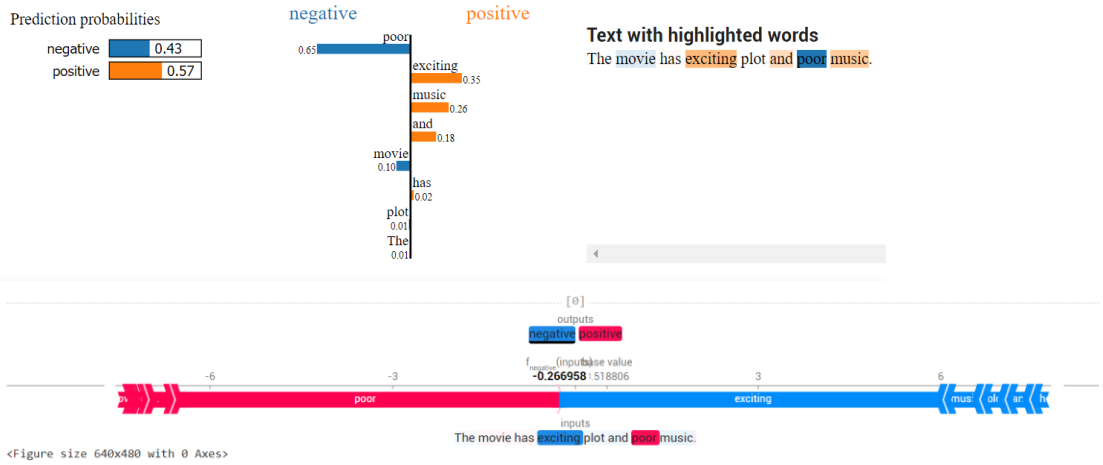
  TA_model_2 是採用 BertModel，並且 embedding 成 512 features。

- 模型表現比較

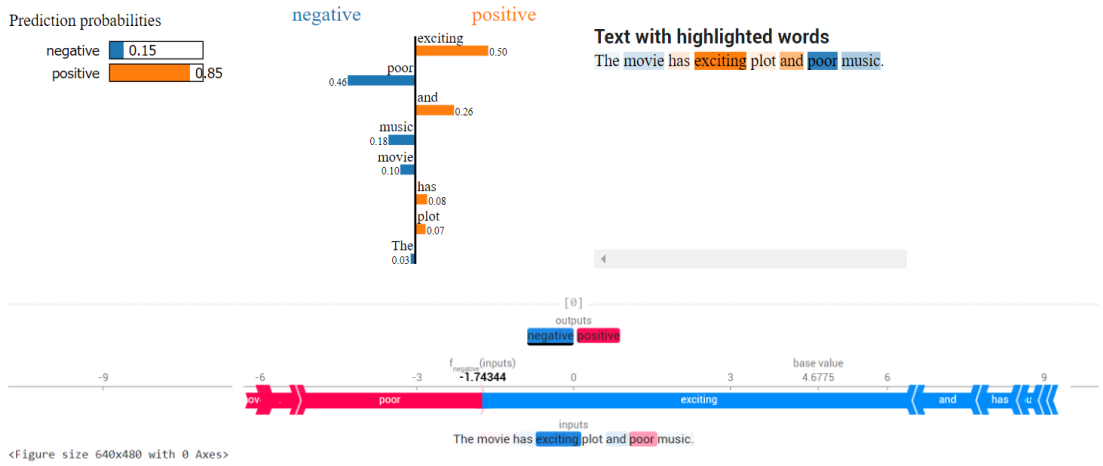
  我用了四句結合正向形容詞與負面形容詞的句子來檢測兩者的差異。

  1. The movie has exciting plot and poor music.

     Model1



     Model2



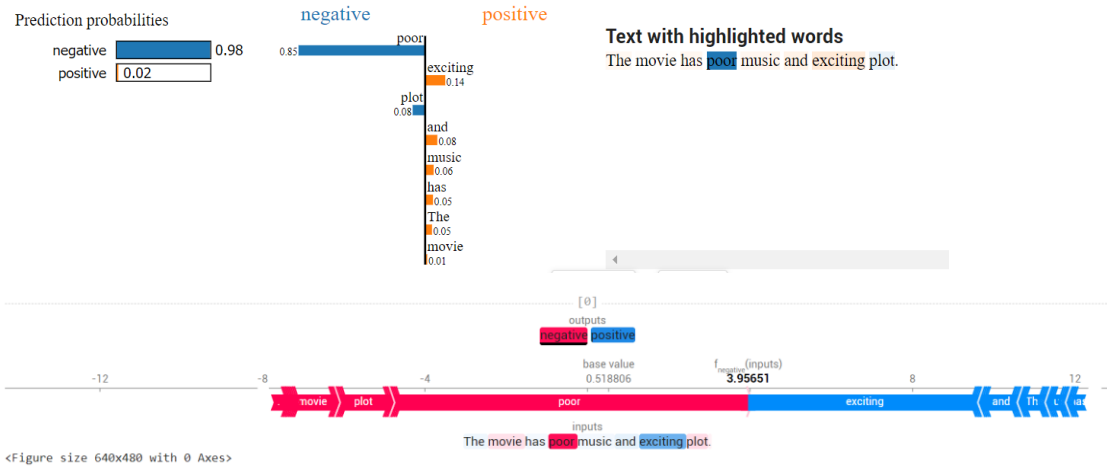
     發現：

     在先正向再負向的句子中，model1 認為這句話是中立的，而 model2 認為是 positive。

     我覺得有兩種可能原因，一是在 model1 中 poor 的負面程度遠大於 model2 覺得的；二是 model2 比較容易受字彙擺放的位置影響他的判斷，所以擺在後面的 poor 就稍顯沒那麼重要。

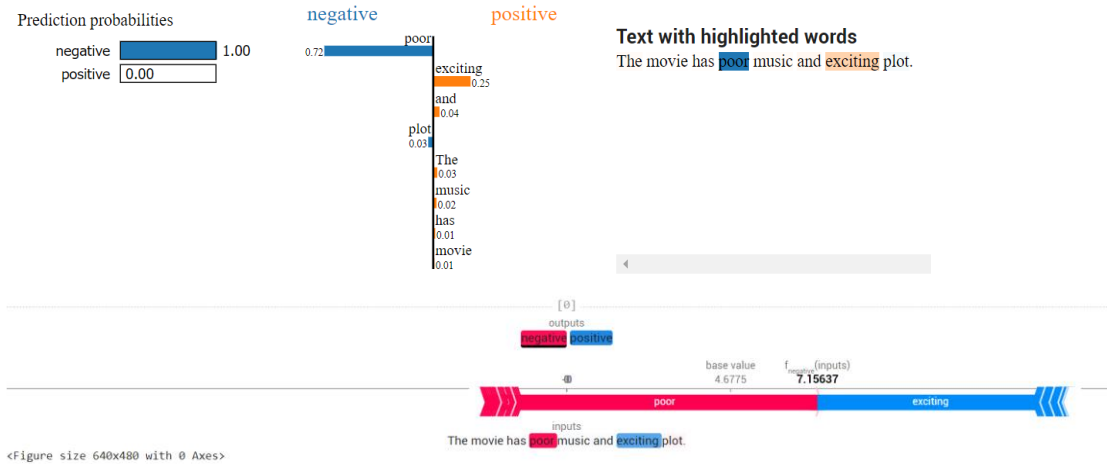
2. The movie has poor music and exciting plot.

Model1



<Figure size 640x480 with 0 Axes>

Model2

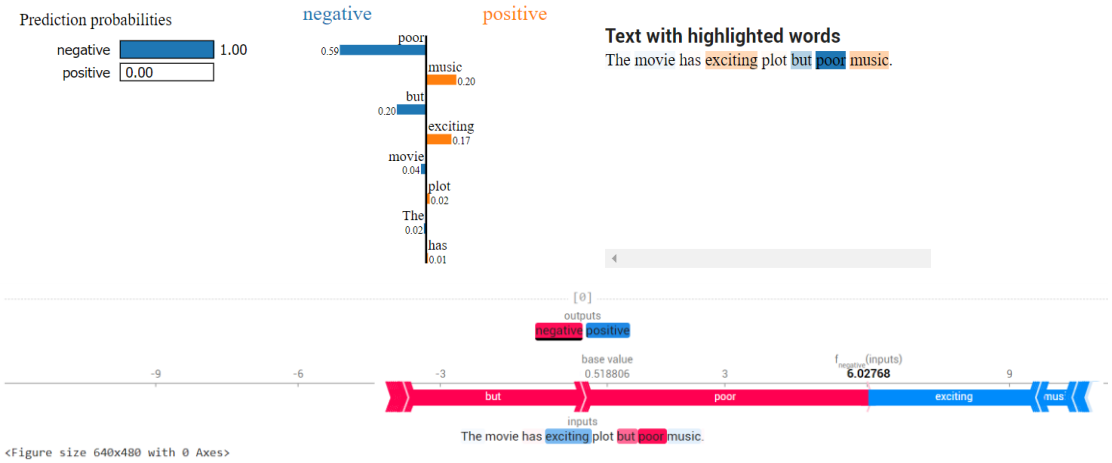


<Figure size 640x480 with 0 Axes>

發現：在先負向再正向的句子中，兩個 model 都認為是 negative。與上一個例子比較可以察覺，當 exciting 擺到後面時，兩個 model 都覺得他沒那麼重要了。

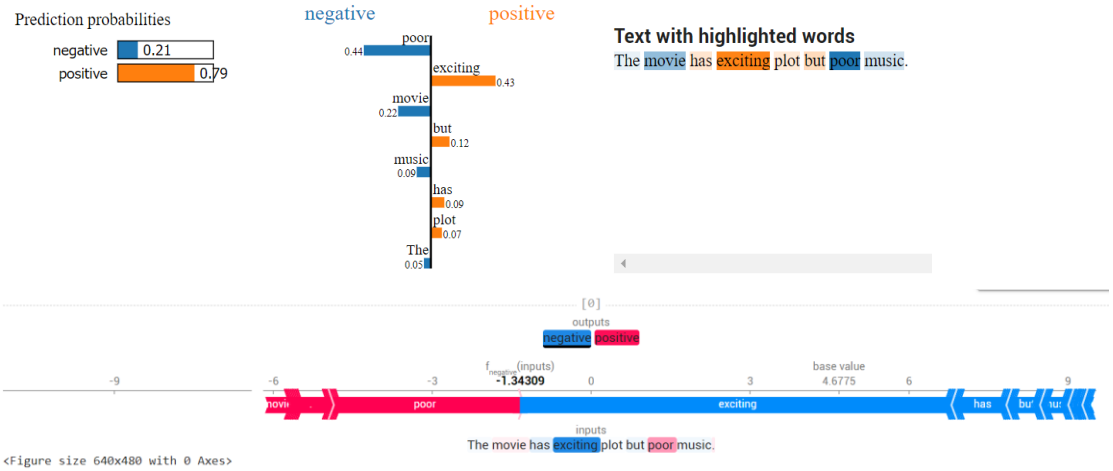
3. The movie has exciting plot but poor music.

Model1



<Figure size 640x480 with 0 Axes>

Model2

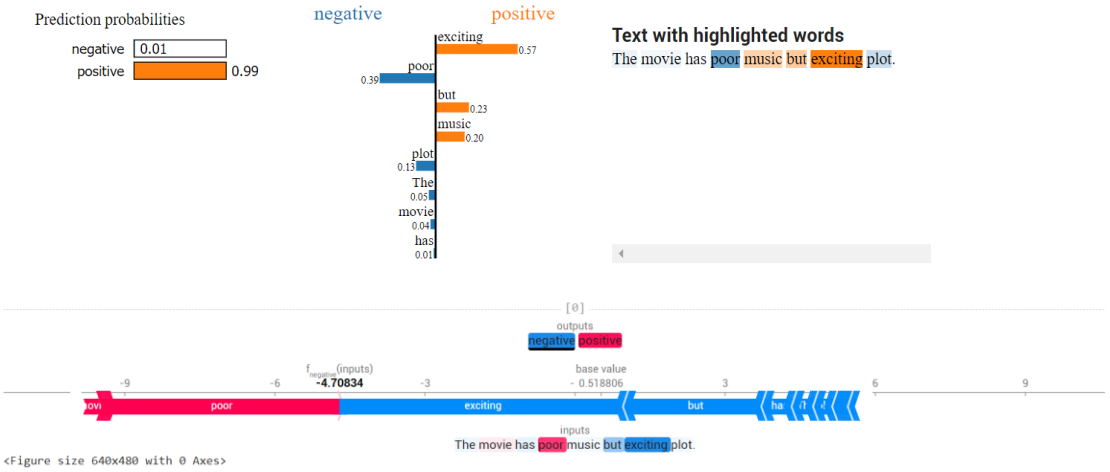


<Figure size 640x480 with 0 Axes>

發現：在先講優點，再用 but 說出缺點的句子中，model1 認為 but 是跟後面的形容詞同向，加重了負面的感覺，因此認為是 negative。而 model2 首先是重視擺在前面的 exciting，再來他認為 but 是與前面的形容詞一致，加強 positive 的比例。

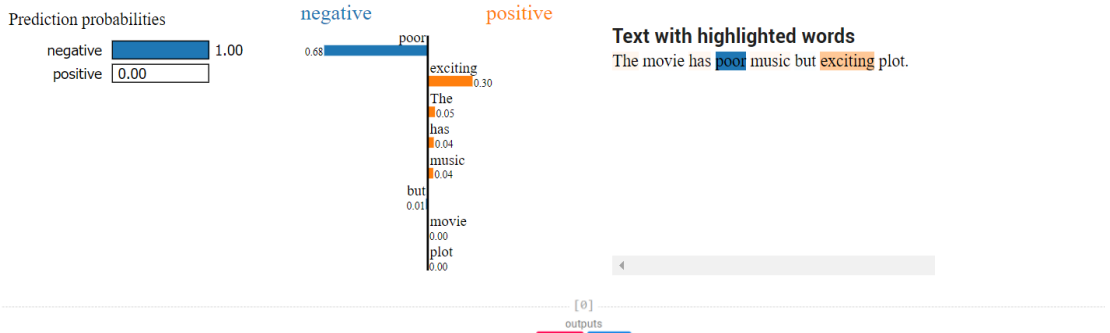
4. The movie has poor music but exciting plot.

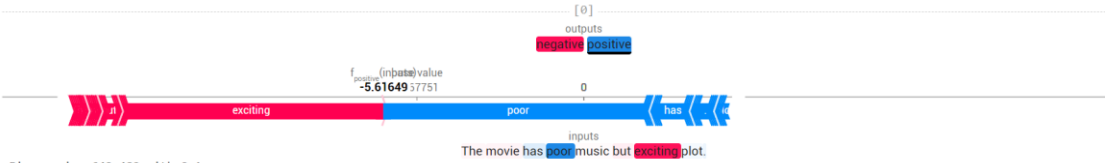
Model1



<Figure size 640x480 with 0 Axes>

Model2





<Figure size 640x480 with 0 Axes>

發現：在先講缺點，再用 but 說出優點的句子中， model1 跟上一個例子類似，認為 but 是跟後面的 exciting 同向，所以是 positive；而 model2 跟第二個例子類似，是因為 poor 擺在前面，就認為他比較重要，因此結論是 negative。

- 總比較

  先講的形容詞對結果影響較大，尤其 model2 受的影響更深。

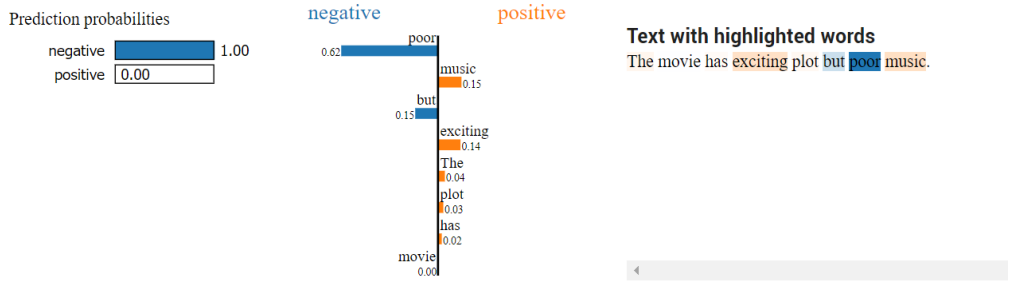
  But 對於 model1 有影響力，他會加強 but 後面的詞對於情緒分類的占比；但是 but 卻對 model2 沒甚麼影響。

3. Compare the explanation of LIME and SHAP.

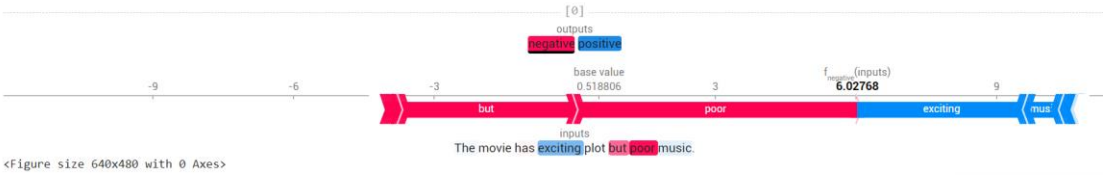
由前一題的例子可以看到，其實多數情況兩者差異不大。

以下使用 model1 預測“The movie has exciting plot but poor music.”為例。

- LIME



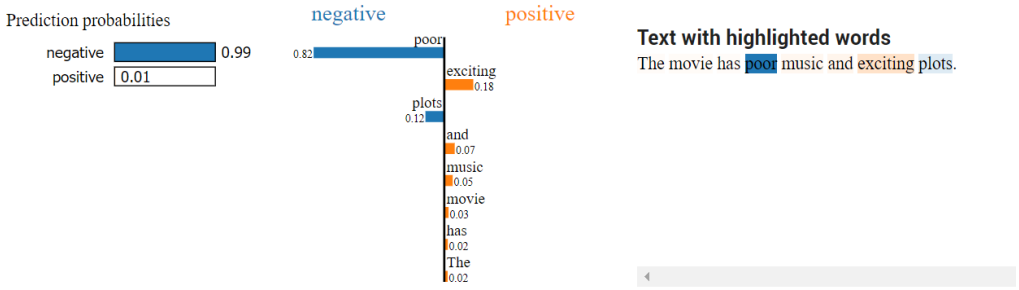
- SHAP



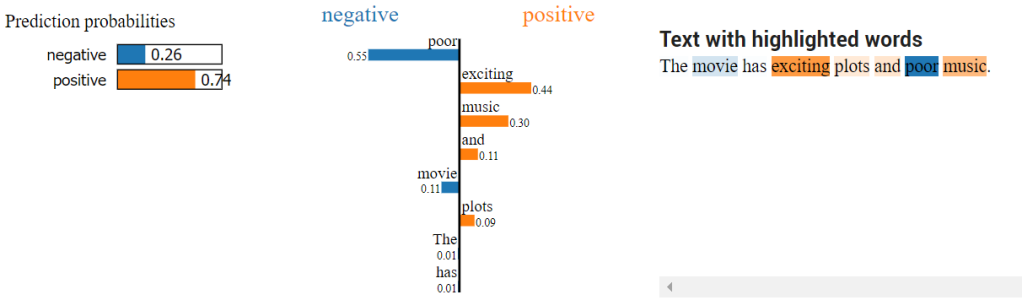
發現：從圖表或是句子的著色明暗可以看出來，在 SHAP 的分析中特別著重帶有情緒的形容詞 poor 與 exciting 以及轉折詞 but，其他佔比不大；但在 LIME 中雖然 poor 佔比很大，但是 music 佔的比重卻與 exciting 差不多。而我認為在此案例中，我認為 SHAP 解釋得更好些，這表示 LIME 使用線性逼近局部模型會有些偏誤。

4. **Try 3 different input sentences for attacks. Also, describe your findings and how to prevent the attack if you retrain the model in the future.**

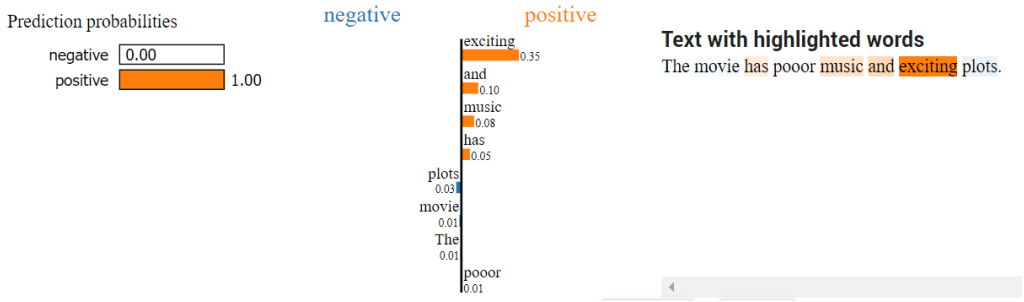
- 原句：The movie has poor music and exciting plots.
  結果：negative 99%



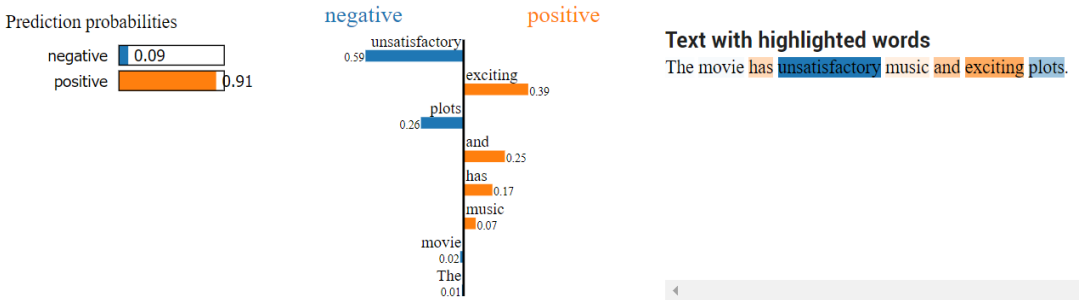
- 攻擊一：改變列順序 The movie has exciting plots and poor music.
  結果：positive 74%



- 攻擊二：插入字母 The movie has pooor music and exciting plots.
  結果：positive 100%



- 攻擊三：改成同義詞 The movie has unsatisfactory music and exciting plots.
  結果：positive 91%



- 發現：
  1. 排序越前面的詞彙 model 比較重視，之後 train 時應該要注意，如果是用 and 連接的話，前後的比重應該一致。
  2. model 沒辦法偵測是不是手誤多打了一個字母。之後或許可以先偵測輸入是否有拼字錯誤。
  3. 改成同義詞後，model 判斷字彙的強烈程度不同，結果就差很多。應該要建一個同義詞的字典，讓同義詞有相同的重要性。