AI HW4 REPORT

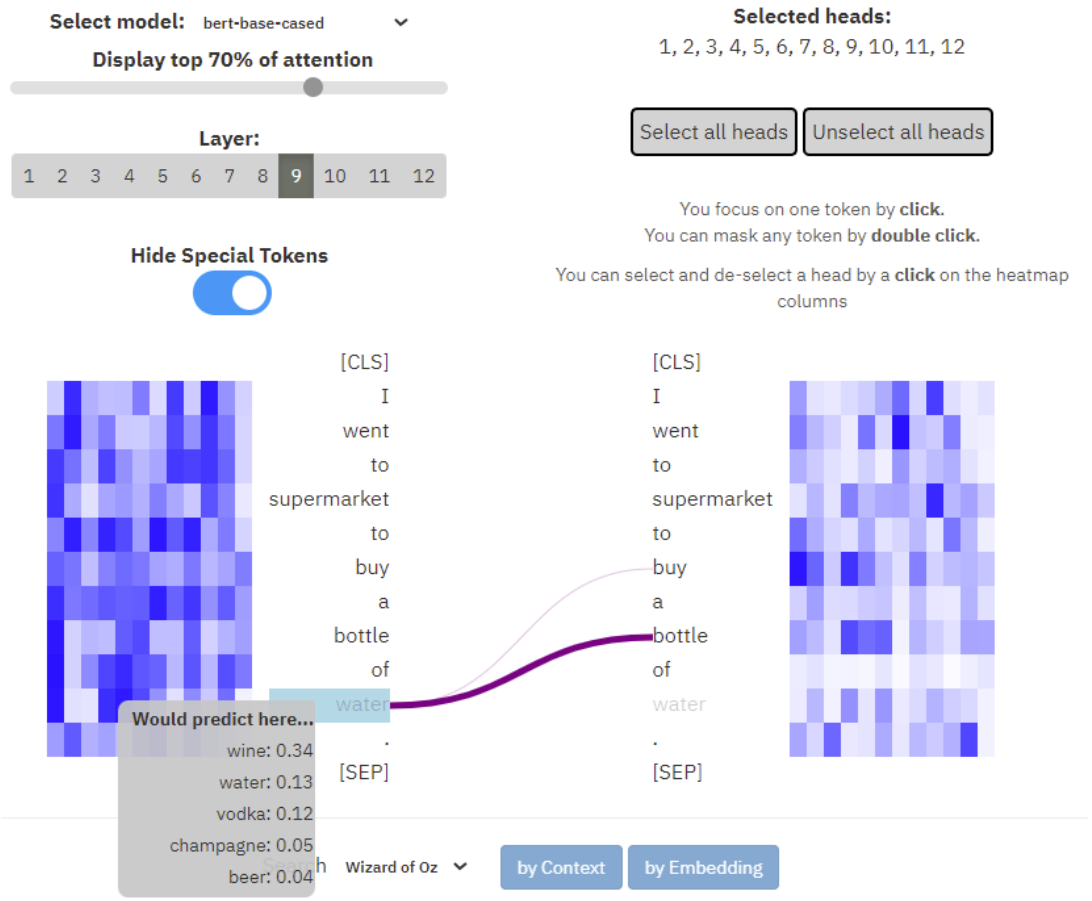
109550167 陳唯文

1. Attention Mechanism

Bert 使用的自注意力機制 (Self-attention mechanism)簡單來說就是讓一個句子的每個單字對前後文的不同單詞分配不同權重，而這使它跟我們在 HW2 練習的 n-gram model 有了重大的區別，因為同一個詞彙在現實生活中會在不同情況有不同含意，這時的語意判斷就需要上下文的輔助。而 Bert 在經過訓練後會使每一個單詞間產生不同關係的神經網路，使它能更接近實際上會發生的情況。

Input sentence :"I went to supermarket to buy a bottle of water."

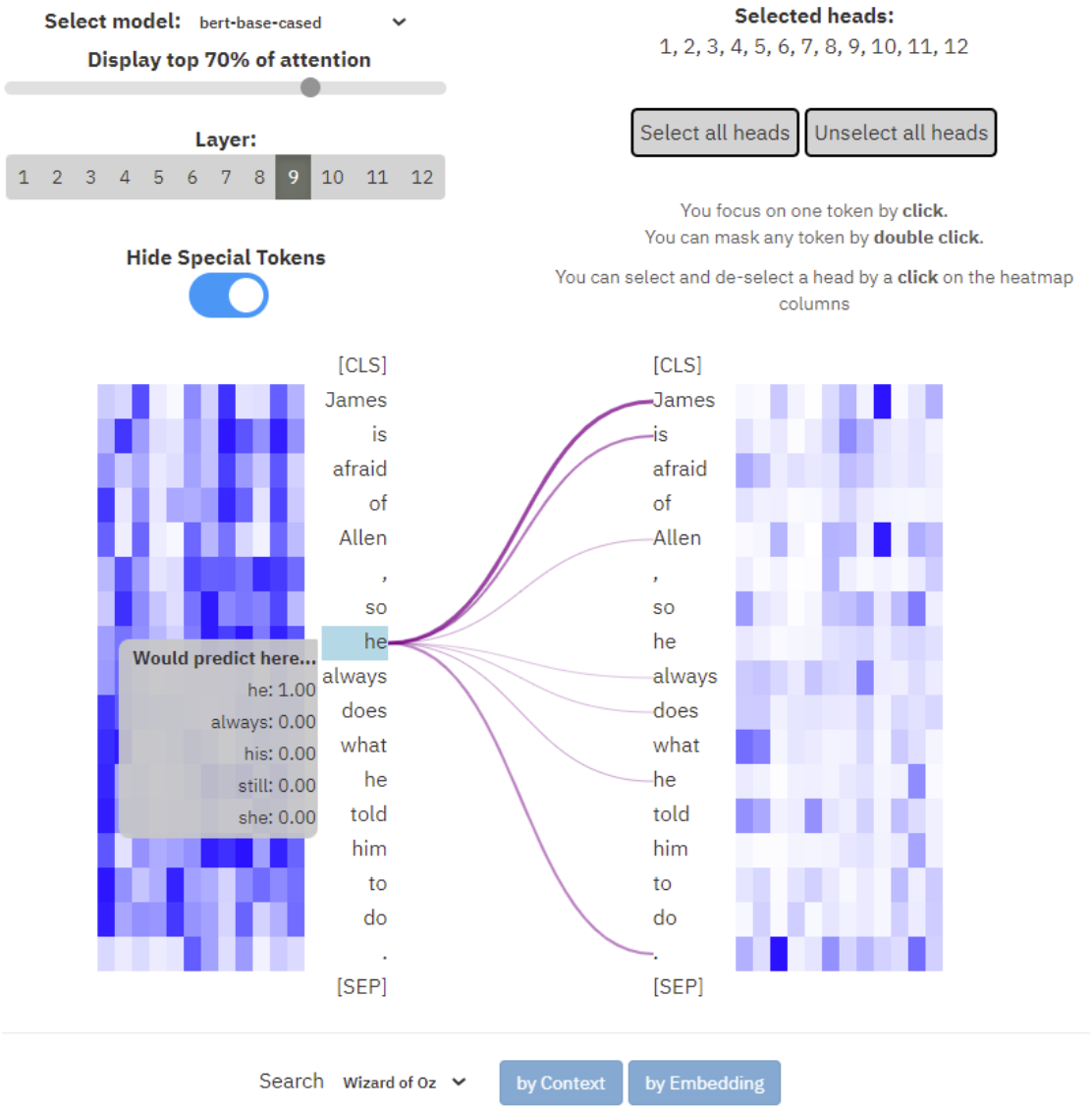
在 exBert 中可以 mask token，所以在這裡我把 water 這個單字 mask 掉，並觀察 layer 9 的結果。可以發現 exBert 將 attention 幾乎都放在了單位量詞'bottle'上，這也使得它能成功預測出可能會出現的單字，其中就包括了 water.

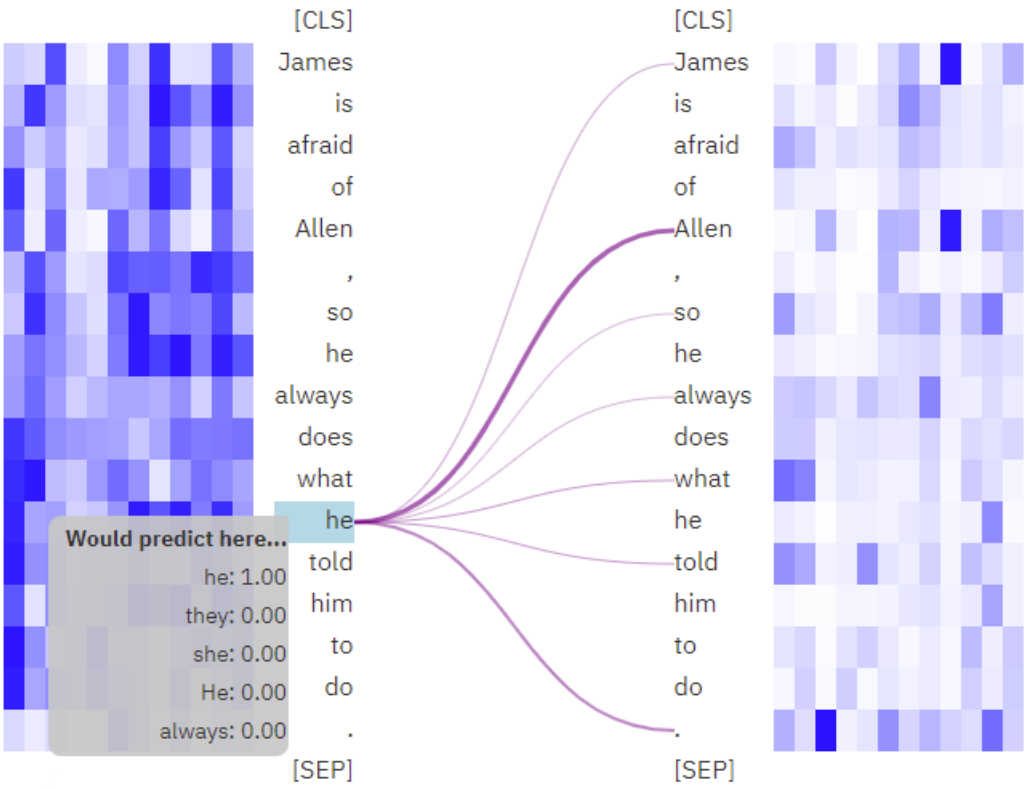


Input sentence :" James is afraid of Allen, so he always does what he told him to do."

在這個句子中我使用了兩個人名和代名詞 he，以觀察 exBert 能否指出正確的對象。從下面的兩個圖中可以發現，不管是第一個 he 還是第二個 he，exBert 正確的將 attention 分別放在 James 和 Allen 上。

要能判斷出這樣的結果，除了需要先把前後文作連結，還需要對文字的一定邏輯思考能力。所以這個不僅代表 Bert 能判斷出不同代名詞所指稱的對象，更代表它已經能夠理解同樣單字在不同情況下的涵義，和這種較複雜的文法結構。





從上面兩個使用 exBert 觀察 attention 的結果來看，Bert 透過 attention mechanism 機制，使用蘊含上下文 / 語境資訊的詞向量，再經過大量文本訓練後，已經能夠研讀出一個句子的正確語意和不同單詞的使用時機，且較 n-gram 這種只看前後單詞的 model 大幅接近一般人類對於詞語運用的水準。

2.Comparison between 2 models

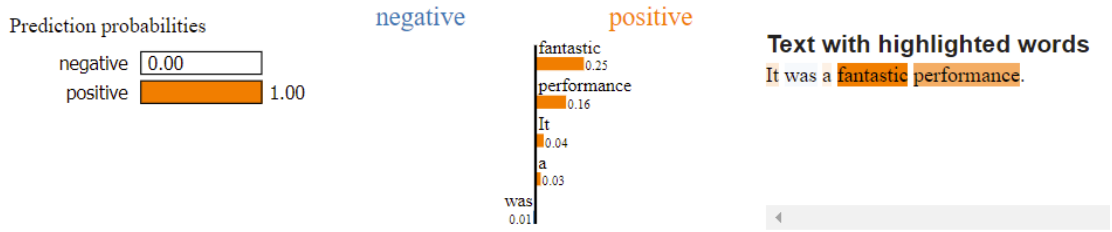
Model used: TA_model_1.pt and TA_model_2.pt

為了比較這兩個不同 model 經過 explainer 後的差異，我一開始先使用了助教提供的兩個明顯表達情感的例句

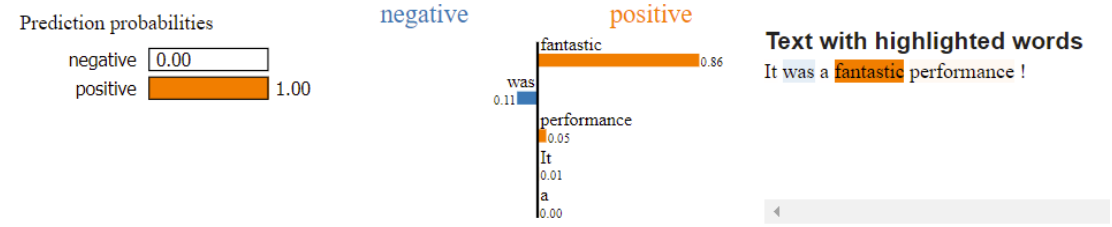
1. It was a fantastic performance !
2. That is a terrible movie.

結果:

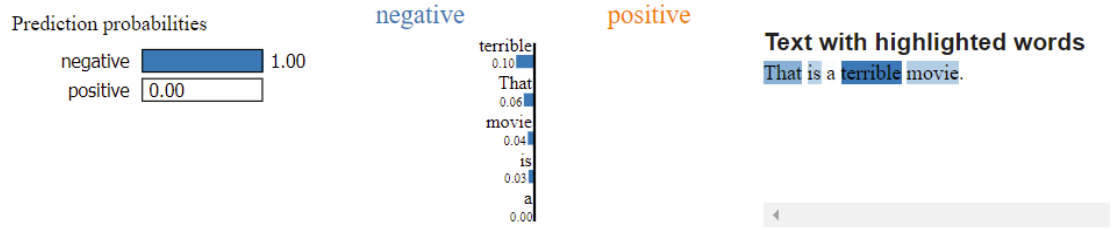
distilbert-base-uncased: sentence 1



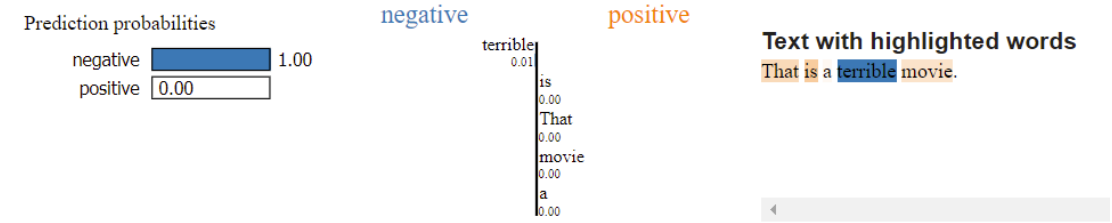
prajjwal1/bert-small: sentence 1



distilbert-base-uncased: sentence 2



prajjwal1/bert-small: sentence 2



從 sentence 1 的解釋結果可以發現，這兩個模型都將 fantastic 作為影響最大的參數，在 bert-small 內它的影響機率更到了 86%。而對 sentence 2，雖然兩個模型內的 terrible 皆是影響最大的參數，但可以發現數值都並不高，bert-small 內甚至只有 0.01。從上面的結果可以推導出在較短的句子內不同模型的表現並沒有太大的差異，但 bert-small 的結果間的浮動會更大。

接著我用 IMDB dataset 的 review 來測試兩個模型對一段文字的預測

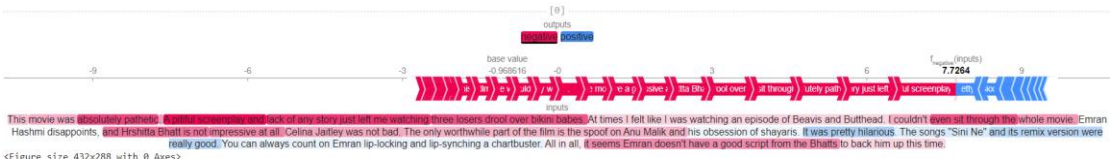
1. This movie was absolutely pathetic. A pitiful screenplay and la
   ck of any story just left me watching three losers drool over b
   ikini babes. At times I felt like I was watching an episode of
   Beavis and Butthead. I couldn't even sit through the whole movi
   e. Emran Hashmi disappoints, and Hrshitta Bhatt is not impressi
   ve at all. Celina Jaitley was not bad. The only worthwhile part

of the film is the spoof on Anu Malik and his obsession of sha yaris. It was pretty hilarious. The songs "Sini Ne" and its rem ix version were really good. You can always count on Emran lip-locking and lip-
synching a chartbuster. All in all, it seems Emran doesn't have a good script from the Bhatts to back him up this time.

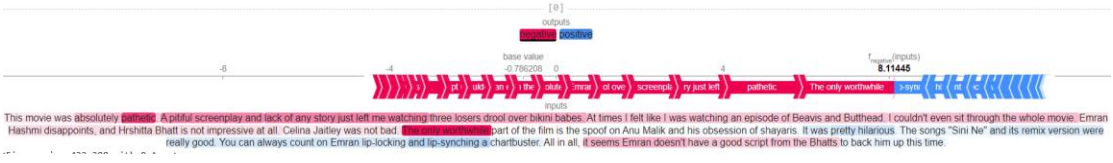
2. Having seen 'only' about 200 Hong Kong films in my time, I have to say this film is among my very top favorites. Not only is t he plot engaging (and in some ways surprising, which these days is rare for any movie), but the chemistry between the two lead actors is superb. Top notch casting! And while often even the most serious HK films tend to insert quite a bit of humor in be tween all the drama and action, often spoiling the mood a bit, here the jokes are kept subtle and woven into the plot, even im proving character relations. The music is also very well done, and the two main themes are very beautiful. With the release of the HK special Edition, they've even cleaned the picture (firs t release was grainy) and the subtitles, even if the quality of the translation is still lacking (nothing new there). All in a ll, if you have to see a HK film that isn't directed by John Wo o or have Chow Yun Fat in it, this should be at least on your s hort list! A truly fascinating and entertaining watch!

結果:

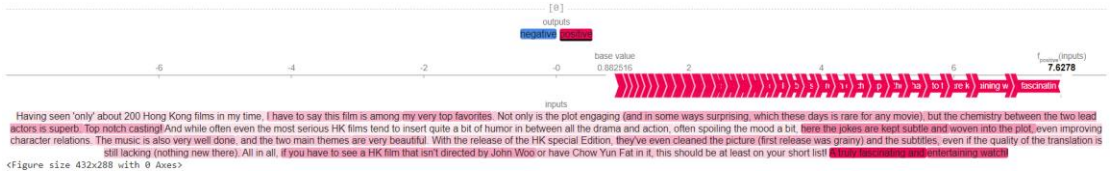
distilbert-base-uncased: review 1



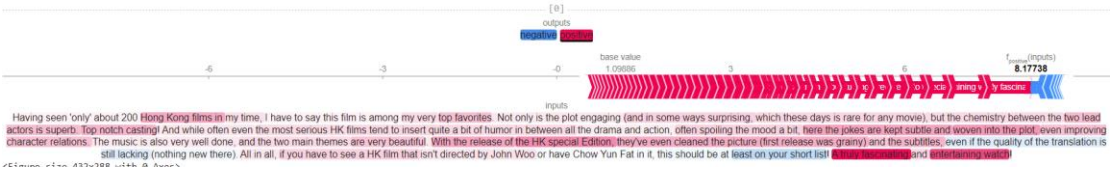
prajjwal1/bert-small: review1



distilbert-base-uncased: review 2



prajjwal1/bert-small: review2



在使用較長的文字來測試後，這兩個模型的差異便開始出現了。儘管它們聚焦的重點是差不多的，但 DistilBert 能更明顯的顯示出哪些是更重要的關鍵，像是第 1 個 review 的 couldn't even sit through 這幾個字，在 DistilBert 內將它歸類為有重要的貢獻度，但 bert-small 內它的貢獻度便不那麼明顯，而實際上它卻是判斷這段 review 的重點之一。而且對於 review 2，它也有判斷錯誤的情況，例如 at least on your short list 這幾個字，bert-small 將他列為會造成 review 為 negative 的項目，但這應該是一個加分項才對，相較之下，DistilBert 便沒有出現一些明顯的錯誤。

會有這樣的結果可能是因為 bert-small 是比起已經簡化過的 DistilBert 更小的 pre-trained Bert variants，所以在簡單句子內它們雖然沒有多少差異，但一旦文字變多，變得更複雜，彼此間的效能差異便會慢慢開始顯現。

3. Explanation Techniques: LIME and SHAP

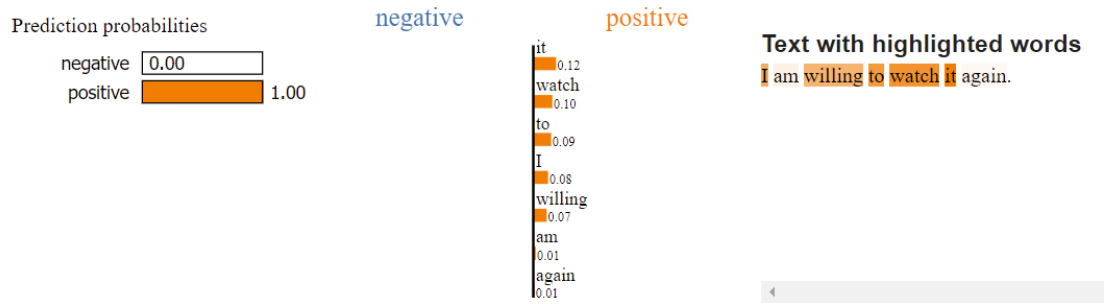
這部分使用的都是 TA_model_1.pt

一開始我先用兩個簡短的句子比較它們的表現

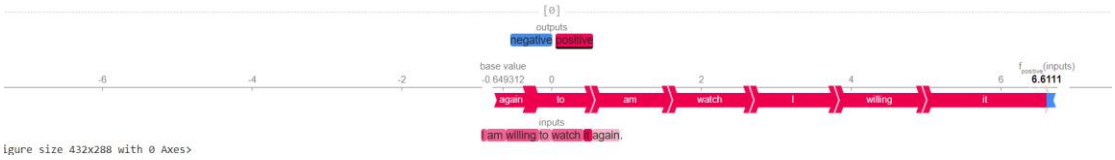
1. 'I am willing to watch it again'
2. 'I would rather watch my 5 years old kid playing sand than watch this movie again.'

而得到的結果如下

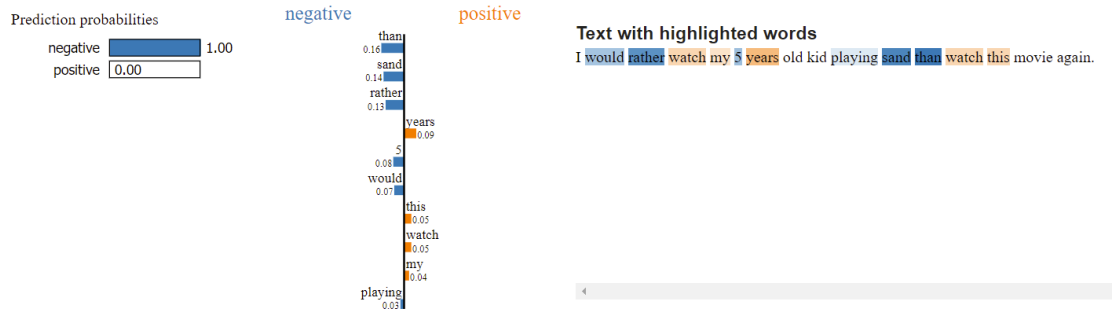
LIME: sentence 1



SHAP: sentence 1



LIME: sentence 2



SHAP: sentence 2

從第一個句子可以發現不管是 LIME 或者是 SHAP，他們認為對 positive 影響最大的單字都是 it，關於這點我並不知道準確的原因，因為實際上在這個句子 it 只是一個無足輕重的代名詞而已，我認為這可能跟我們使用的 model 有關。而這一句最重要表達 positive 的單字 willing，在 SHAP 內它的貢獻度為第二，但在 LIME 內它的影響卻是排到相當後面。所以對於這一個句子，SHAP 應該是更符合我們解釋一句話的方式。
至於第二個句子，雖然並沒有包含能夠表達情感的詞彙，但這兩個 explainer 都能成功的解釋 rather 跟 than 便是影響這句語意的關鍵點。

接下來我使用 IMDB dataset 的一個 positive review 跟一個 negative review 來測試 LIME 跟 SHAP 的解釋性。

1. Positive

“Some films just simply should not be remade. This is one of them. In and of itself it is not a bad film. But it fails to capture the flavor and the terror of the 1963 film of the same title. Liam Neeson was excellent as he always is, and most of the cast holds up, with the exception of Owen Wilson, who just did not bring th

e right feel to the character of Luke. But the major fault with t
his version is that it strayed too far from the Shirley Jackson s
tory in it's attempts to be grandiose and lost some of the thrill
of the earlier film in a trade off for snazzier special effects.
Again I will say that in and of itself it is not a bad film. But
you will enjoy the friction of terror in the older version much
more."

2. Negative

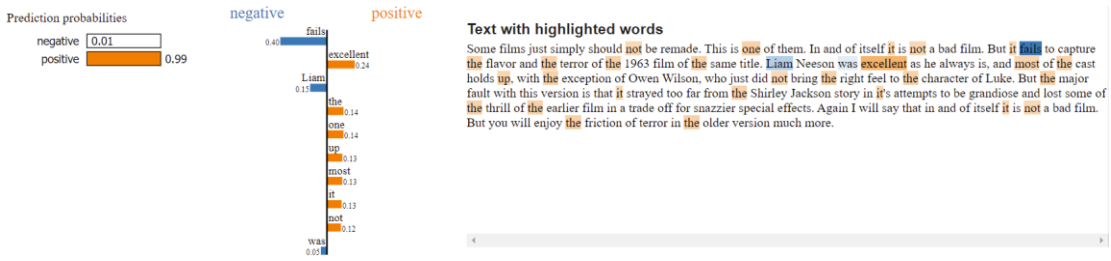
"No one can say I wasn't warned as I have read the reviews (both user & external), but like most of us attracted to horror movies... curiosity got this cat. (Come on, we all scream at the people in the movie not to go into the dark room, but you know that's horror aficionados are always dying to know what's in there even if we know it'll be bad).

The bottom line is that this movie left me angry. Not because it pretends to be real (who cares...gimmicks are allowed), or because the actors and dialogue are so lame (is this an unusual event in horror movies?) or even because the movie is so bad (and I am being polite here). What really got me mad is that the film is not only a rip off of BWP, but also a half-hearted lazy rip off at that.

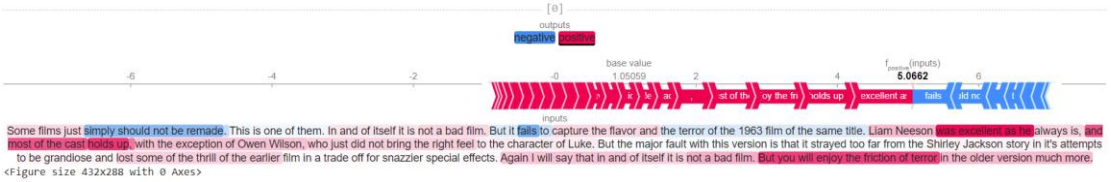
I don't believe in sacred cows and if they thought they could outdo BWP then kudos to them, but they didn't even try. The movie was made with little effort or care and that is the most unforgivable sin in horror (or any) movie!"

結果:

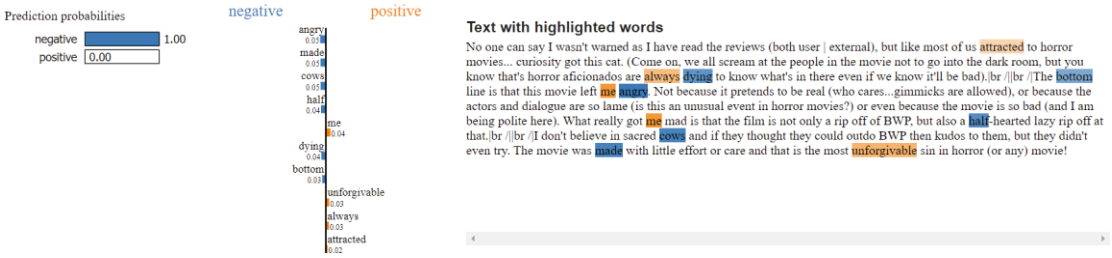
LIME: review 1



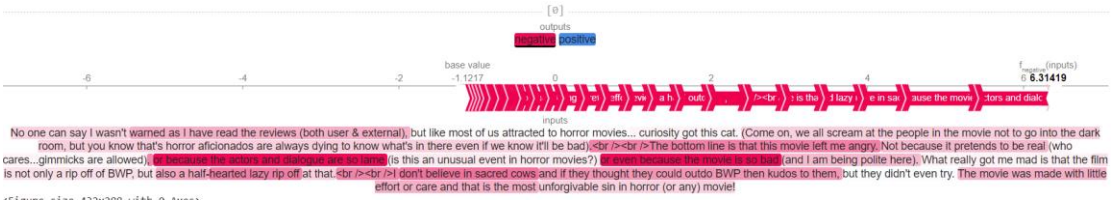
SHAP: review1



LIME: review 2



SHAP: review2



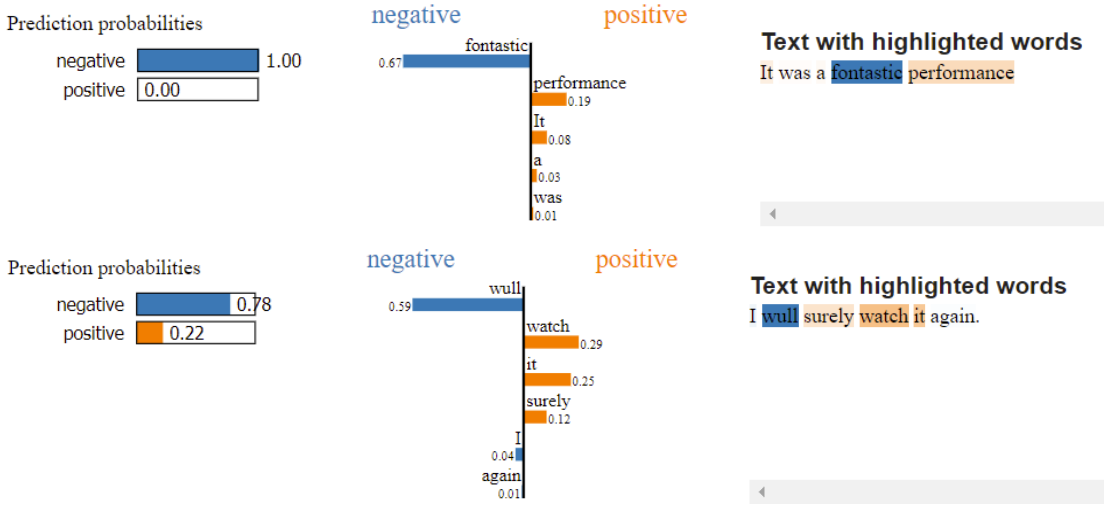
從上面的結果可以發現，相較於 LIME 只把重點聚焦在單個詞彙上，SHAP 的解釋方法則是用數個單字所組成的一小段有意義的文字來提供他們的貢獻度，而很顯然 SHAP 的解釋能更讓我們了解這段話被歸類到 positive 或 negative 的真正原因。

例如第一個 review 的結果，儘管 LIME 有標示出 excellent 是使 review 機率為 positive 的重要原因，但接下來的幾個單字: Liam, the, one 都是沒什麼意義的詞彙，而 Liam 這個人名甚至會使 review 變為 negative 提高，這樣的解釋很難讓人真正了解將一段話判斷為 positive 或 negative 的依據。而 SHAP 卻是標示出了幾個重要的關鍵，像是 was excellent as, But you will enjoy the friction of terror，甚至是片語 holds up，這些有意義的一段文字比起一個單字，更能說服我們為什麼 Bert 會將一個 review 判斷為正面或負面。

而會有這樣的結果可能是因為 LIME 為了能夠給出簡單明瞭的解釋，使用的是 local linear approximation，與一個 model 實際上分類的方式並不是完全一致的，但判斷一段上百字文章的情感並不只是一個單字就能夠給出原因，還需要上下文的依據、詞向量等複雜並難以表達的原因，這就不是 LIME 能夠解釋的範圍。而 SHAP 則是對 feature 計算 Shapley value，並依照貢獻度的高低給出結果，這樣的解釋方式更符合我們在理解一段話會使用的判斷標準。

4. Attack

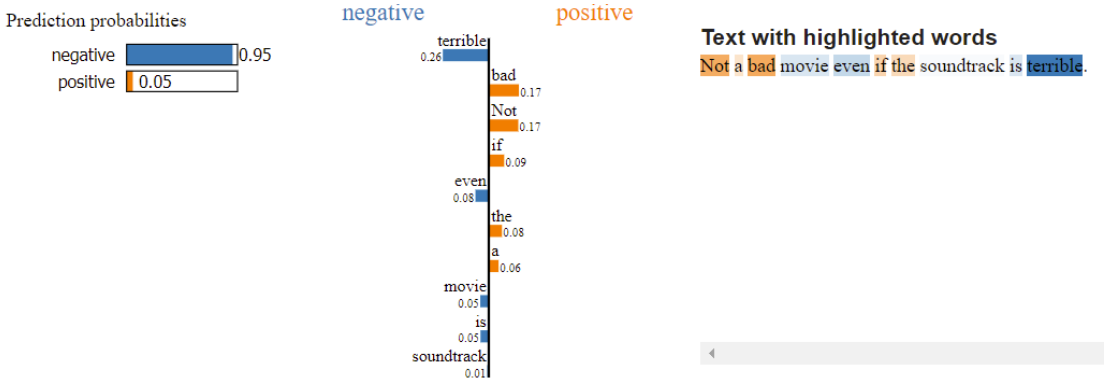
1. Misspelling



在第一張圖我將 fantastic 替換成 fontastic，儘管只差了一個字母，卻讓原本 1.00 positive 的評論變成 1.00 negative。第二張圖，我則是將 will 替換成 wull，結果一樣使得原本為 1.00 positive 的評論轉為 negative。由此可知，儘管可能只有出現一點點的改變，依舊會使 model 出現截然不同的結果。

2. Ambiguous sentence

這裡我嘗試了一些可能會讓 model 判斷錯誤的組合



Input sentence :"Not a bad movie even if the soundtrack is terrible."
這一句話的真正意義是在說這部電影並不差，但出現的結果卻是 bert 被最後一個單字 terrible 影響，而判斷這句話為 negative.

Input sentence :"The performance of the main actor is neither good nor bad."

在這個句子裡並沒有表達明確的 positive 或 negative，但卻因為 neither 跟 nor 使 bert 判斷它為 100% negative。在這一部份我嘗試的兩個句子都沒有出現不合理的文字組合，而是實際上可能會出現的例子，但 DistilBert 卻都沒有辦法成功辨別它們的情緒。

從上面嘗試出的四種結果來看，不論是簡單的拼字錯誤或是一些較困難的語意轉折，DistilBert 都有出錯的可能性，可能是因為 DistilBert 和實際上的 Bert 性能還是有一定的差距，或是在 train 的時候 model 並沒有接觸過這類型 data 的經驗，如果是這樣，可以在訓練的階段增加這些可能會讓 model 出現錯誤的 dataset，儘管不能完全解決問題，但應該能提升一部份的正確率。

5. Problem

雖然這次的 lab 並沒有要求我們寫 code，但還是可以學到滿多的東西，比較困難的部分應該就是要先閱讀關於 LIME 跟 SHAP 的相關資料，才能理解和比較這些工具出現的結果和原因，而在各種測試的過程也能讓人思考出現問題背後的成因，更進一步了解 bert 和可以再加強的部分。