Introduction to Artificial Intelligence -HW4

109350008　　張詠哲

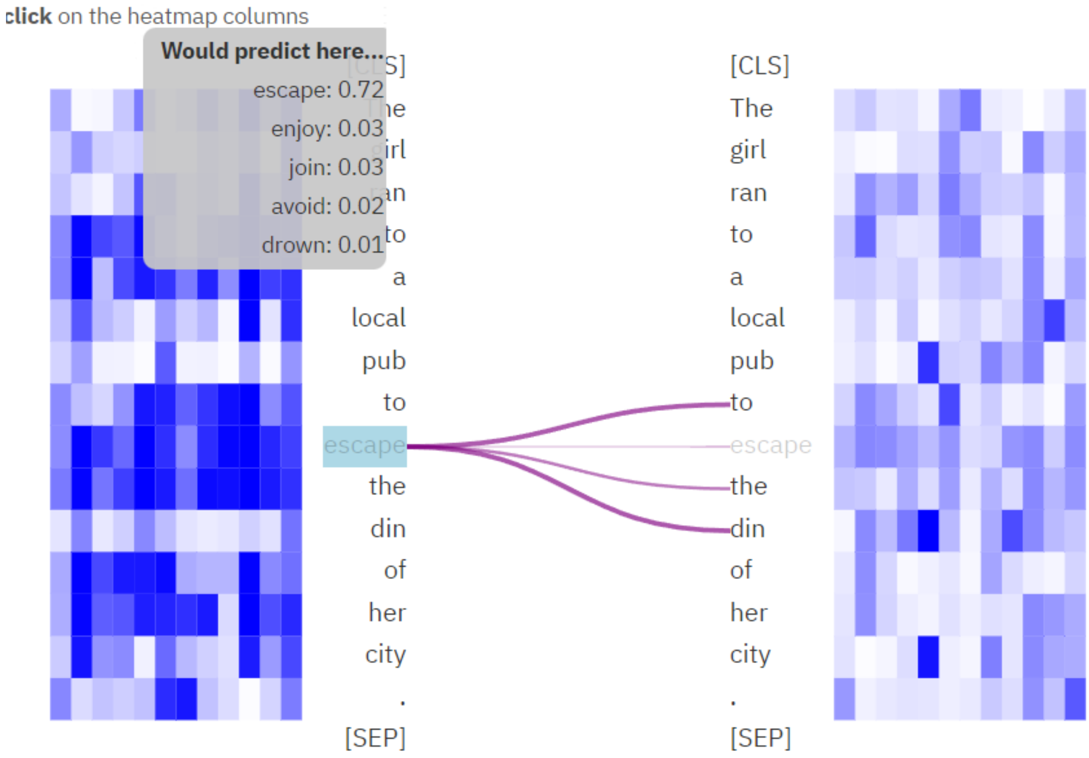
1. Describe your understanding and findings about the attention mechanism by exBERT.

BERT（Bidirectional Encoder Representation for Transformers）是一種基於 Transformers 的 NLP 模型。 BERT 在兩項任務上進行了預訓練: Masked Language Model（MLM）和 Next Sentence Prediction（NSP）。MLM 要對輸入的文本進行隨機 masking，將部分單詞替換為 [MASK] 標記。然後模型需要通過上下文來預測被遮住的單詞是什麼。而 NSP 需判斷第 2 個句子在原始文本中是否跟第 1 個句子相接。在這個部分，會使用 ExBERT，一個的可視化 BERT，用來了解 BERT 的 Attention 機制。其中會使用的包括 bert-base-cased 和 distilbert-based-uncased。

1.1 使用 bert-base-cased

- **The girl ran to a local pub to escape the din of her city**

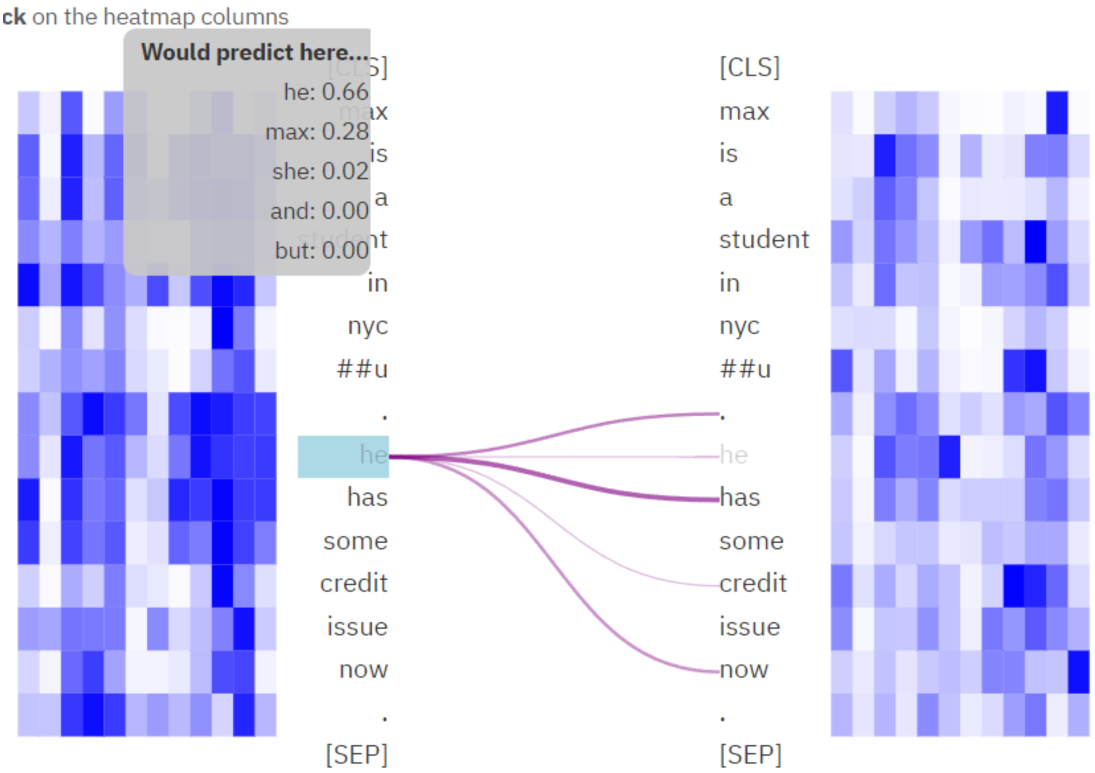
以下是 later = 7，且使用默認的句型: The girl ran to a local pub to escape the din of her city. 並遮住 escape 的結果。



正如上圖所表現，如果遮住 “escape” 這個詞，讓 BERT 預測這裡應該出現哪個詞。在 BERT 的第 7 層中，獲得關注的詞是 ”to” 、”the” 、”din” 、 ”escape”。這些詞正是與”escape” 相關的詞。

- **Max is a student in NYCU. He has some credit issue now.**

接著改為輸入 Max is an student in NYCU. He has some credit issue now. 並把 He 遮住(layer = 12)



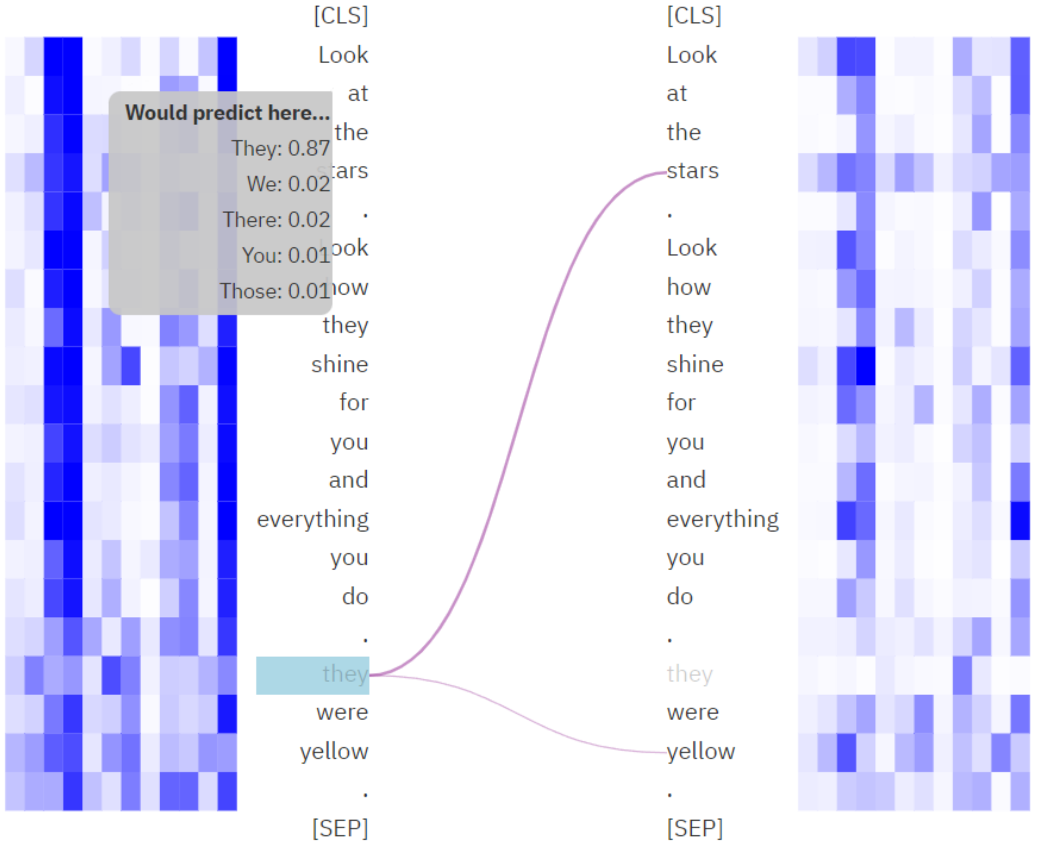
ck on the heatmap columns

Would predict here...
he: 0.66
max: 0.28
she: 0.02
and: 0.00
but: 0.00

這句話中我遮住了主詞，看看 BERT 如何會如何預測或是關注哪個詞。而由上圖可看出， BERT 關注的是第一句話的主詞”Max”。還可以觀察到”He”出現在遮住位置的機率最高。說明了雖然有一些機率預測為是女生，但 BERT 還是覺得 Max 通常是男性的名字。

- **Look at the stars. Look how they shine for you and everything you do. they were yellow..**

再來改為輸入 Look at the stars. Look how they shine for you and everything you do. they were yellow. 並遮住末句的 they 所產生的結果 (layer = 12)。



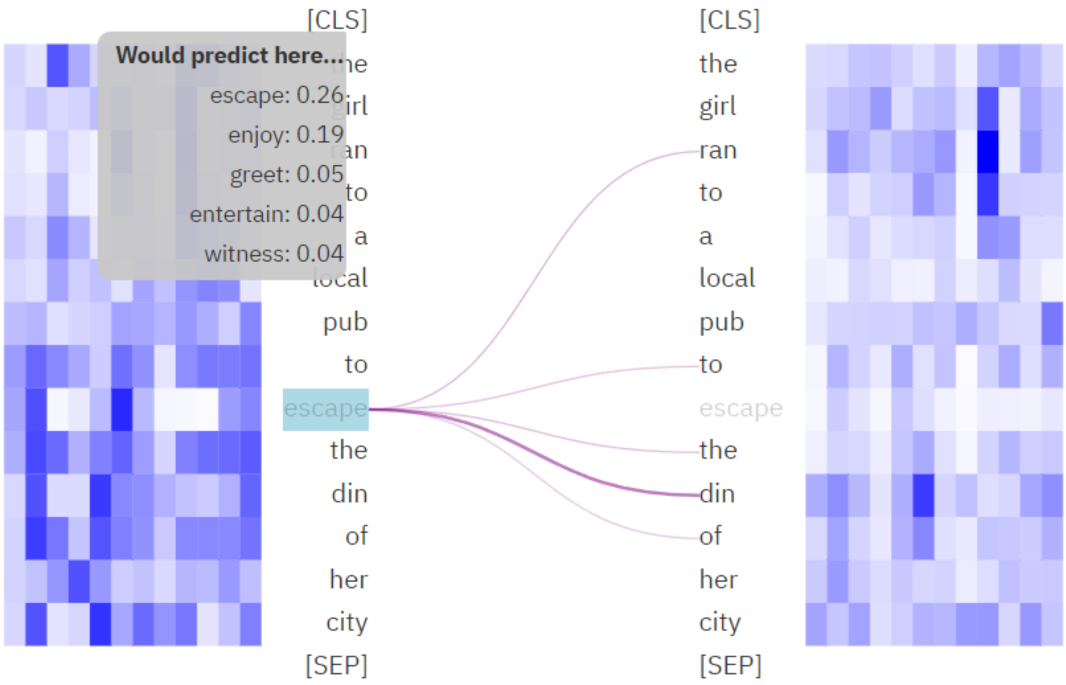
這部分是想了解 BERT 是否能夠通過上下文來預測相關代名詞。而正如在上圖中所看到。BERT 在預測遮住的詞時將注意力集中在”stars”以及”yellow”。且預測出來機率最高的也為”they”而不是 “you” 。這表明 BERT 可以通過分析上下文來預測相關代名詞，並知道應該注意哪個詞。

由上面的測試，大致上了解一些 BERT 的 Attention 機制。相較於 n-gram 模型或其他沒有使用這個機制的模型，Attention 機制會使 BERT 表現更好。

1.2 使用 distilbert-based-uncased

BERT 是一個較大的模型，有很多層(包括 340 兆個參數)。而 DistilBERT 是在 BERT 基礎上進行了簡化(只有 66 兆個參數)，但是所需要的計算資源更少。而在訓練方式也不同，例如 BERT 的訓練過程中包含 MLM 和 NSP。而 DistilBERT 只訓練了 MLM。因此，DistilBERT 的預訓練時間比 BERT 短，同時也較容易實現，但也保持了不錯的表現也提升了模型速度。
我使用與 BERT 相同的三個句子並遮住相同的詞。下面是實驗結果(layer = 6)。

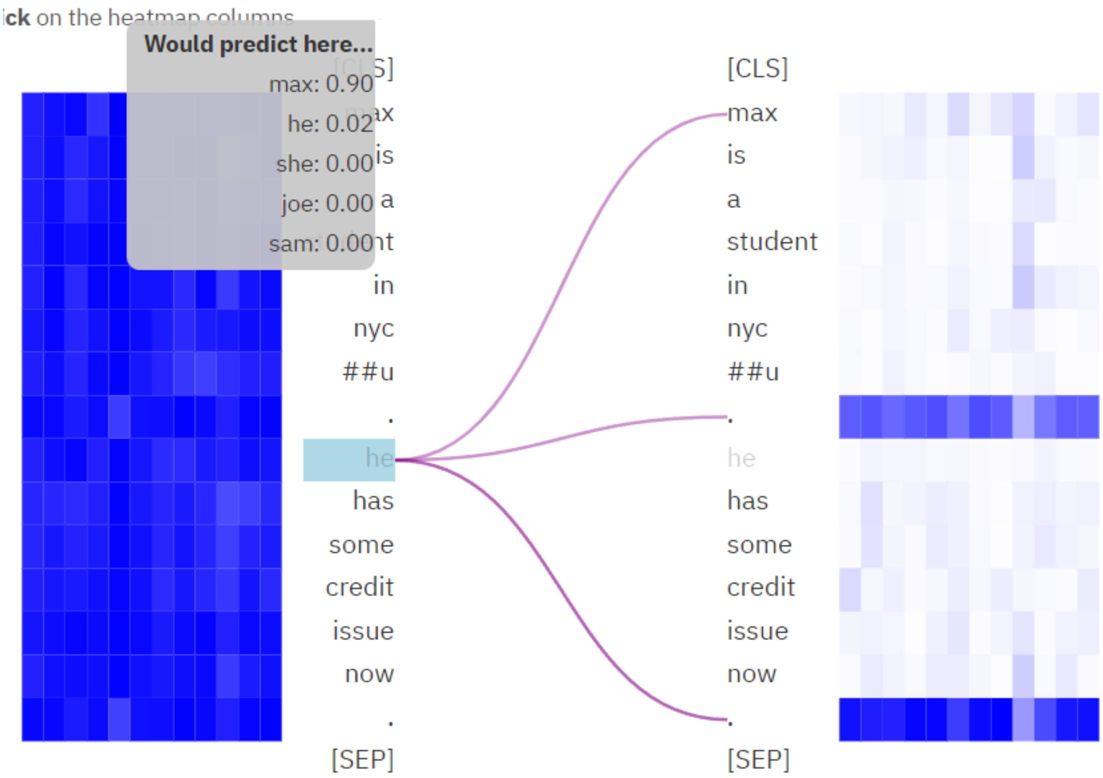
- **The girl ran to a local pub to escape the din of her city**



Would predict here...
escape: 0.26
enjoy: 0.19
greet: 0.05
entertain: 0.04
witness: 0.04

如上圖所示，DistilBERT 也是能夠正確預測，它關注的詞與 BERT 關注的詞相似，多了”ran”、”of ”。 這表明 DistilBERT 從原始的 BERT 中學習了一些基本知識。

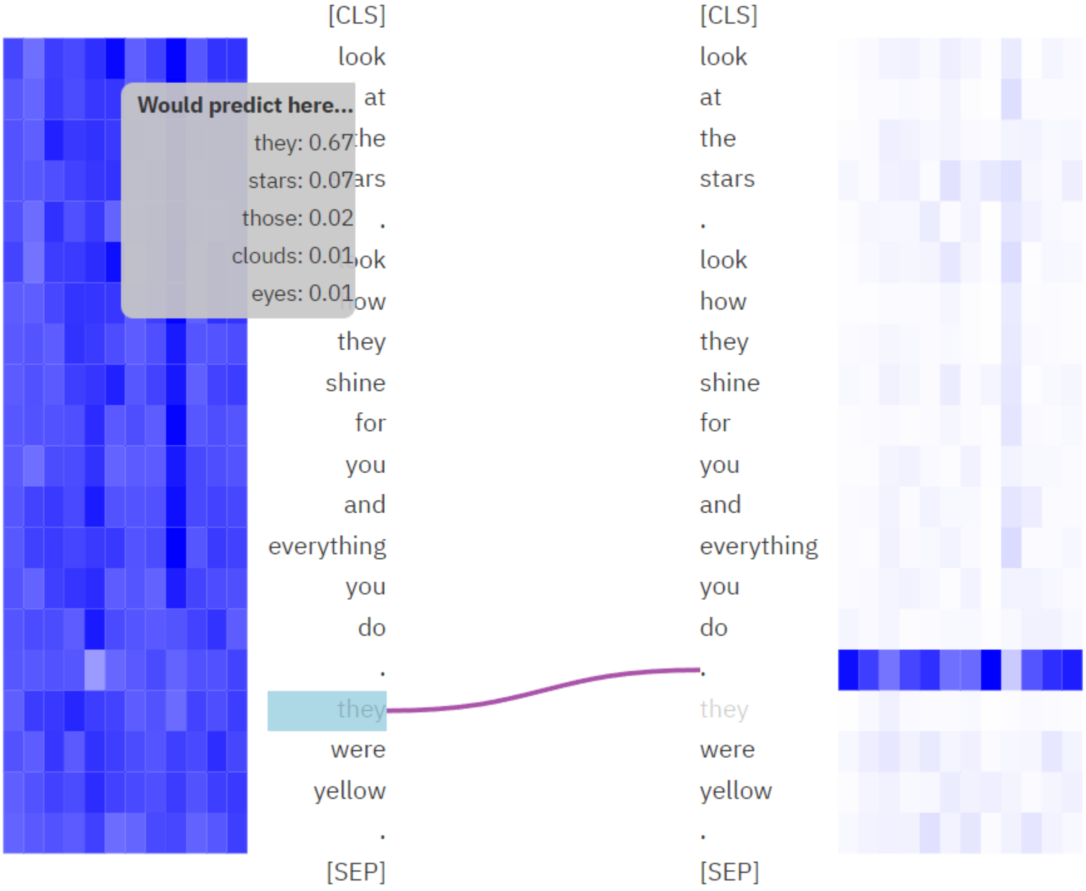
- **Max is a student in NYCU. He has some credit issue now.**



Would predict here...
max: 0.90
he: 0.02
she: 0.00
joe: 0.00
sam: 0.00

由上圖所示，DistilBERT 和 BERT 的預測結果相差很多，它改為直接預測出原本上文的主詞。

- **Look at the stars. Look how they shine for you and everything you do. they were yellow.**



由上圖可看出，相較於 BERT，DistilBERT 也是可以預測出正確的詞，但是機率大大的下降。而且關注位置不知道為何跑到句點上。也跑出一些有點莫名其妙地預測詞(“clouds”、”eyes”)。

1.3 比較 bert-base-cased & distilbert-based-uncased

先比較 Max is a student in NYCU. He has some credit issue now.這句，DistilBERT 可以預測被遮住的詞可能為”Max”或”He”。儘管這兩個都是正確答案並且模型知道 Max 是男性名字。但更好的答案應該是”He” (更接近日常英文用法)。但是 DistilBERT 只有 2% 的概率在”He”。這表明 DistilBERT 的性能相較於原始 BERT 差。而 Look at the stars. Look how they shine for you and everything you do. they were yellow.的結果，DistilBERT 也是可以預測出正確的詞，但是機率大大的下降。而且關注位置不知道為何跑到句點上。也跑出一些有點莫名其妙

地預測詞(“clouds”、”eyes”)。

綜合上面的，雖然 DistilBERT 可以在一些簡單的句子中正確預測單詞，但是如果句子過於復雜，DistilBERT 可能會表現不如 BERT。但是 DistilBERT 的優點就是快且小，對於想練習或是只是做簡單任務就相對友善很多。

2. Compare the explanation of LIME and SHAP. (30%)

Explainable AI 是為了能夠使人們可以理解和解釋機器學習模型的決策過程。傳統上的一些機器學習模型如 DNN 被稱為黑盒模型(black box model)，因為它們難以解釋，使用者無法理解模型如何從輸入數據到輸出預測的過程。而 explainable AI 的目標就是提供關於模型決策的解釋，從而增加對模型的信任度、可解釋性和可靠性。LIME 和 SHAP 都是常用的 explainable AI 的方法:

- **LIME** 是一種模型無關的解釋方法，它通過在局部區域內對模型進行近似，來解釋模型的預測。LIME 將原始輸入數據進行微小的修改和擾動，生成許多新的樣本，然後將這些樣本作為輸入對模型進行預測。通過觀察這些新樣本的預測結果，LIME 可以生成一個簡單且可解釋的模型，該模型解釋了原始模型在該局部區域內的行為。

- **SHAP** 則是一種基於博弈論的解釋方法，它基於 Shapley 值的概念，用於衡量每個特徵對於預測結果的貢獻。SHAP 通過計算特徵值的各種組合，來確定每個特徵的重要性(Shapley 值)，該值表示特徵在所有可能的特徵組合中對預測結果的平均貢獻。通過將各個特徵的 Shapley 值相加，我們可以獲得對於整個模型預測的解釋。以下為 Shapley 值的公式:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!}[v(S \cup \{i\}) - v(S)]$$

首先會使用一些簡單的句子，接著會採用作業二中的 IMDB 資料集來做比較。而其中也會使用跟比較助教提供的 TA_model_1.pt 以及 TA_model_2.pt。

- **簡單句子**

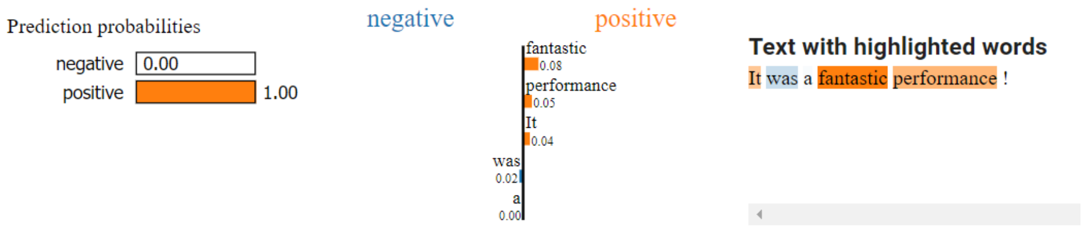
所使用的句子如下:

1. short 1: 'It was a fantastic performance !'
2. short 2: 'That is a terrible movie.'
3. short 3: 'This movie was a complete disappointment, lacking compelling storytelling and engaging performances.'

- **LIME**

short 1

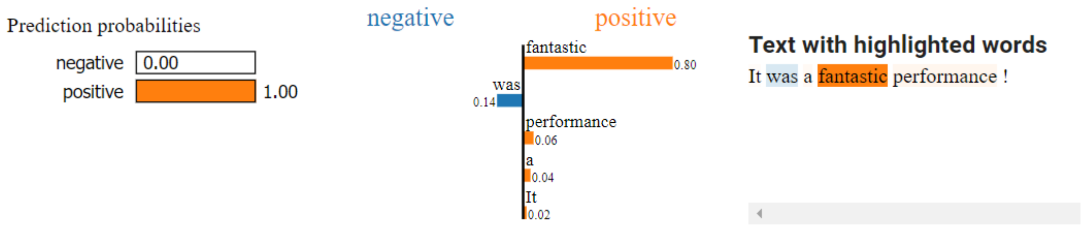
```
model 1:
```



Text with highlighted words

It was a fantastic performance !

```
model 2:
```

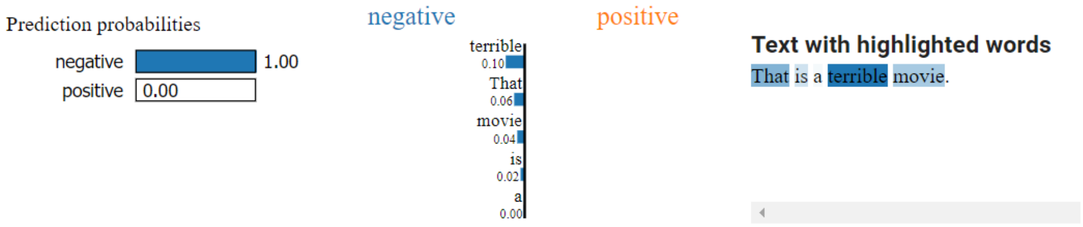


Text with highlighted words

It was a fantastic performance !

short 2

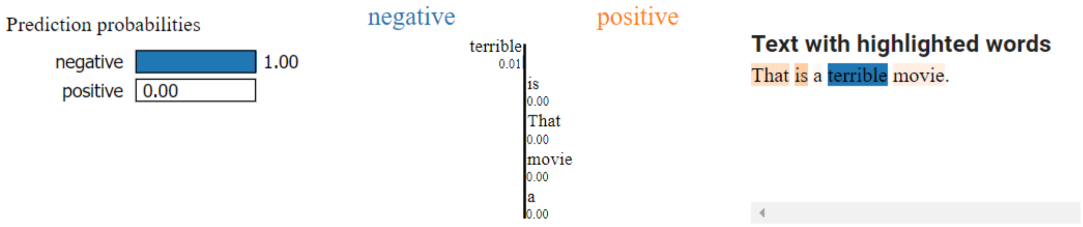
```
model 1:
```



Text with highlighted words

That is a terrible movie.

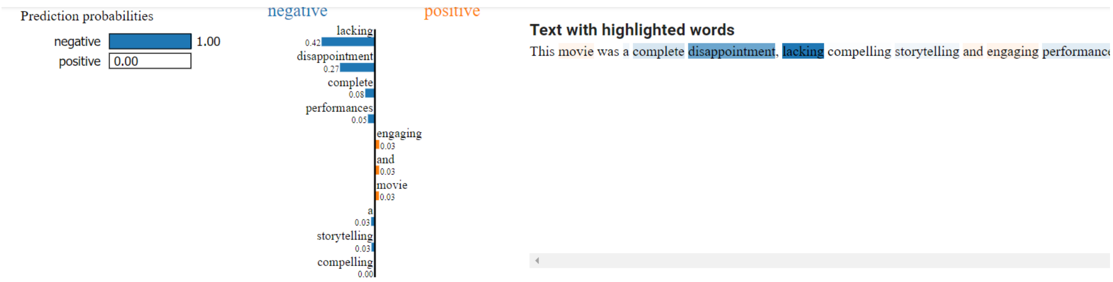
```
model 2:
```



Text with highlighted words

That is a terrible movie.

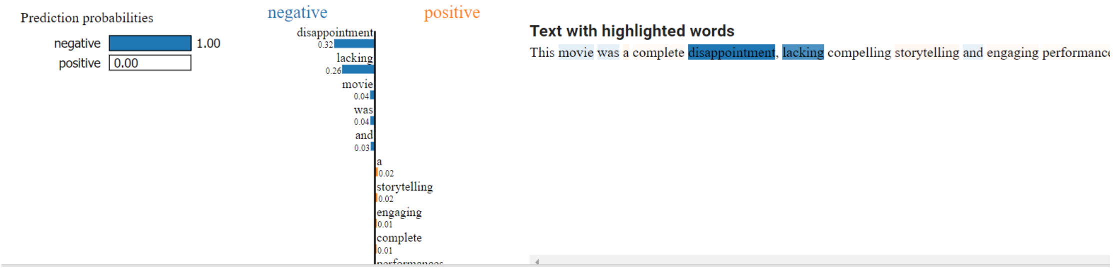
short 3



Text with highlighted words

This movie was a complete disappointment, lacking compelling storytelling and engaging performanc

```
model 2:
```



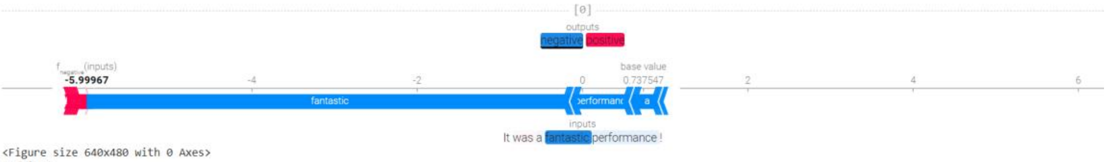
Text with highlighted words

This movie was a complete disappointment, lacking compelling storytelling and engaging performanc

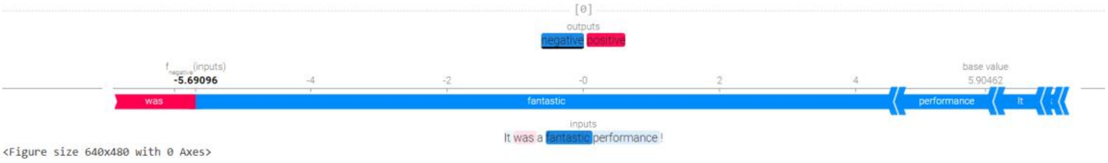
■ SHAP

short 1

```
model 1:
```



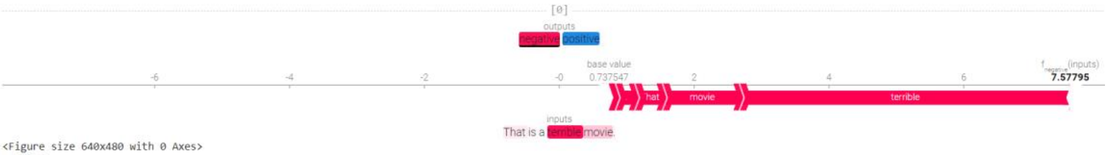
```
<Figure size 640x480 with 0 Axes>
model 2:
```



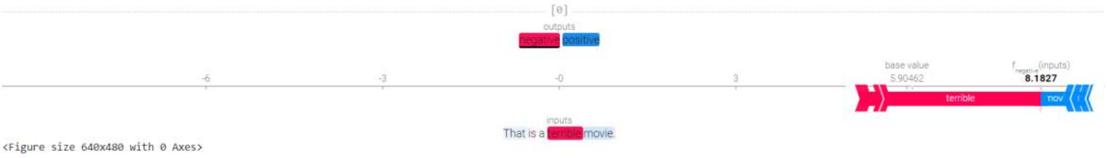
```
<Figure size 640x480 with 0 Axes>
```

Example 2

```
model 1:
```

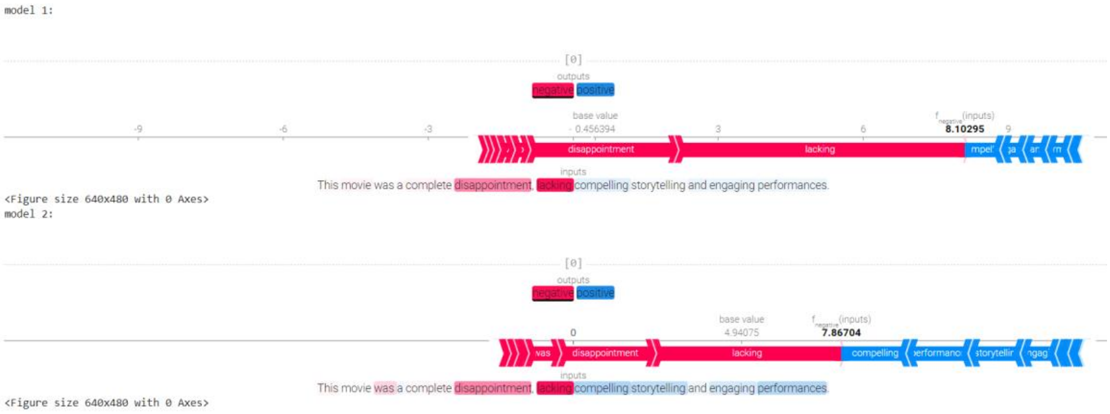


```
<Figure size 640x480 with 0 Axes>
model 2:
```



```
<Figure size 640x480 with 0 Axes>
```

Example 3



```
model 1:
```

This movie was a complete disappointment, lacking compelling storytelling and engaging performances.

```
<Figure size 640x480 with 0 Axes>
model 2:
```

This movie was a complete disappointment, lacking compelling storytelling and engaging performances.

```
<Figure size 640x480 with 0 Axes>
```

- **IMDB**

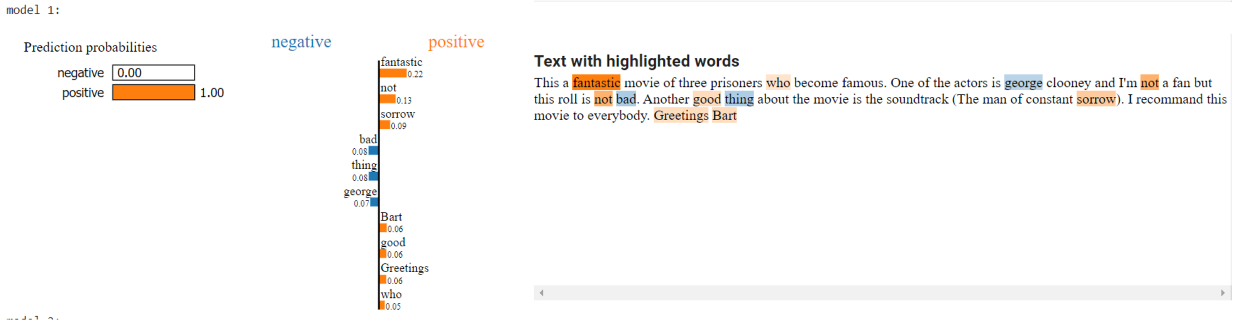
所使用的較長的句子如下:

1. **long1** : "This a fantastic movie of three prisoners who become famous. One of the actors is george clooney and I'm not a fan but this roll is not bad. Another good thing about the movie is the soundtrack (The man of constant sorrow). I recommand this movie to everybody. Greetings Bart"
2. **long2** : "I sure would like to see a resurrection of a up dated Seahunt series with the tech they have today it would bring back the kid excitement in me.I grew up on black and white TV and Seahunt with Gunsmoke were my hero's every week.You have my vote for a comeback of a new sea hunt.We need a change of pace in TV and this would work for a world of under water adventure.Oh by the way thank you for an outlet like this to view many viewpoints about TV and the many movies.So any ole way I believe I've got what I wanna say.Would be nice to read some more plus points about sea hunt.If my rhymes would be 10 lines would you let me submit,or leave me out to be in doubt and have me to quit,If this is so then I must go so lets do it."
3. **long3** : "This show was an amazing, fresh & innovative idea in the 70's when it first aired. The first 7 or 8 years were brilliant, but things dropped off after that. By 1990, the show was not really funny anymore, and it's continued its decline further to the complete waste of time it is today.<br /><br />It's truly disgraceful how far this show has fallen. The writing is painfully bad, the performances are almost as bad - if not for the mildly entertaining respite of the guest-hosts, this show probably wouldn't still be on the air. I find it so hard to believe that the same creator that hand-selected the original cast also chose the band of hacks that followed. How can one recognize such brilliance and then see fit to replace it with such mediocrity?

I felt I must give 2 stars out of respect for the original cast that made this show such a huge success."

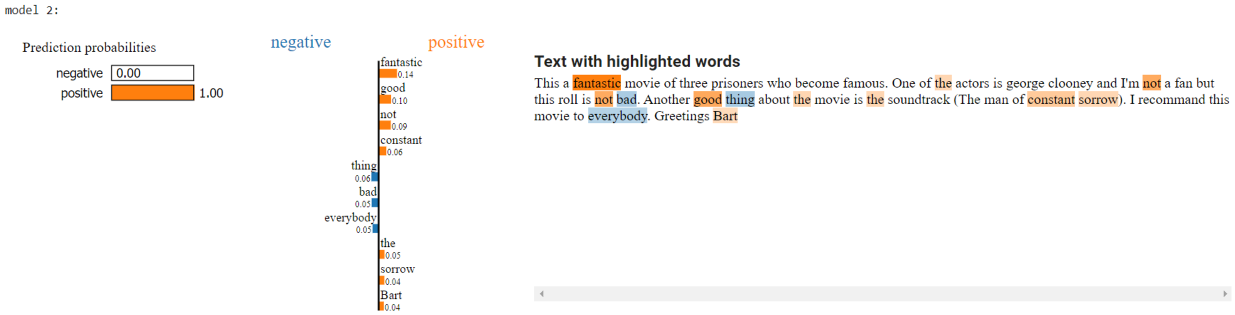
■ LIME

Long 1



Text with highlighted words

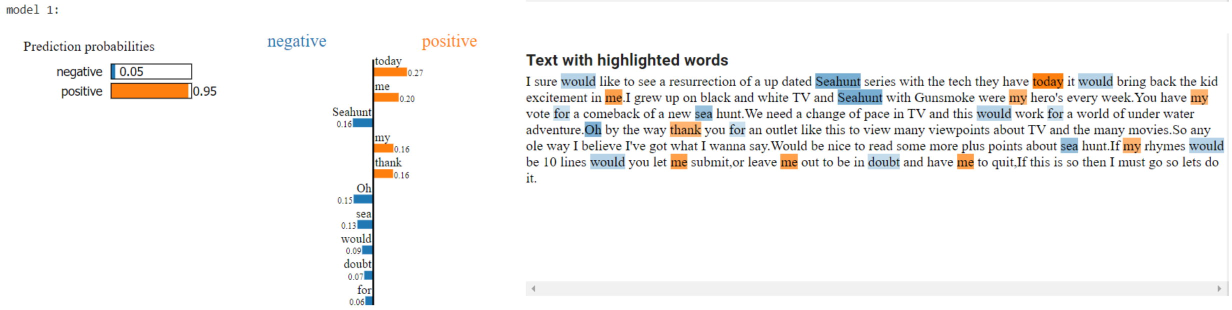
This a fantastic movie of three prisoners who become famous. One of the actors is george clooney and I'm not a fan but this roll is not bad. Another good thing about the movie is the soundtrack (The man of constant sorrow). I recommand this movie to everybody. Greetings Bart



Text with highlighted words

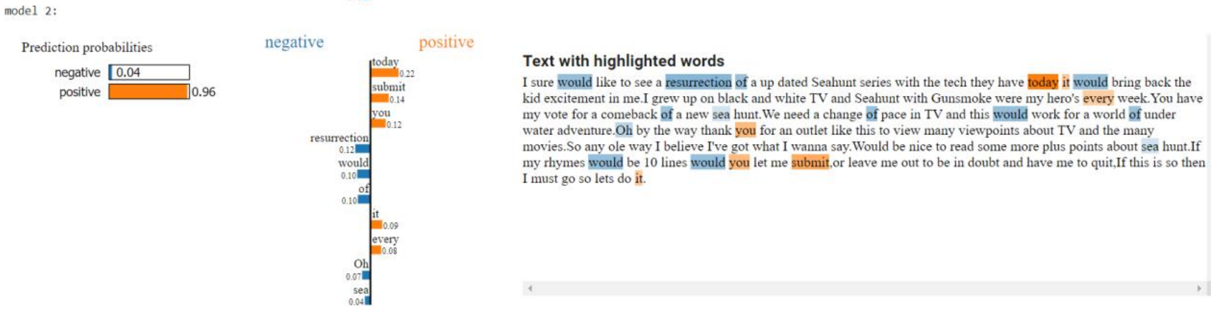
This a fantastic movie of three prisoners who become famous. One of the actors is george clooney and I'm not a fan but this roll is not bad. Another good thing about the movie is the soundtrack (The man of constant sorrow). I recommand this movie to everybody. Greetings Bart

Long 2



Text with highlighted words

I sure would like to see a resurrection of a up dated Seahunt series with the tech they have today it would bring back the kid excitement in me.I grew up on black and white TV and Seahunt with Gunsmoke were my hero's every week.You have my vote for a comeback of a new sea hunt.We need a change of pace in TV and this would work for a world of under water adventure.Oh by the way thank you for an outlet like this to view many viewpoints about TV and the many movies.So any ole way I believe I've got what I wanna say.Would be nice to read some more plus points about sea hunt.If my rhymes would be 10 lines would you let me submit,or leave me out to be in doubt and have me to quit,If this is so then I must go so lets do it.

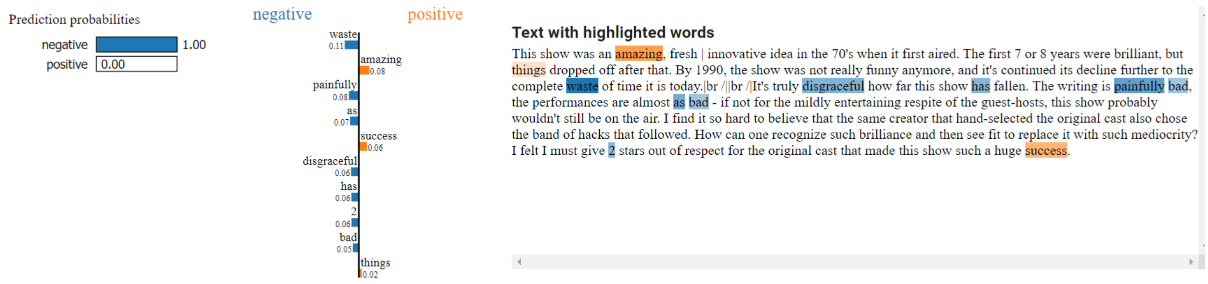


Text with highlighted words

I sure would like to see a resurrection of a up dated Seahunt series with the tech they have today it would bring back the kid excitement in me.I grew up on black and white TV and Seahunt with Gunsmoke were my hero's every week.You have my vote for a comeback of a new sea hunt.We need a change of pace in TV and this would work for a world of under water adventure.Oh by the way thank you for an outlet like this to view many viewpoints about TV and the many movies.So any ole way I believe I've got what I wanna say.Would be nice to read some more plus points about sea hunt.If my rhymes would be 10 lines would you let me submit,or leave me out to be in doubt and have me to quit,If this is so then I must go so lets do it.

Long 3

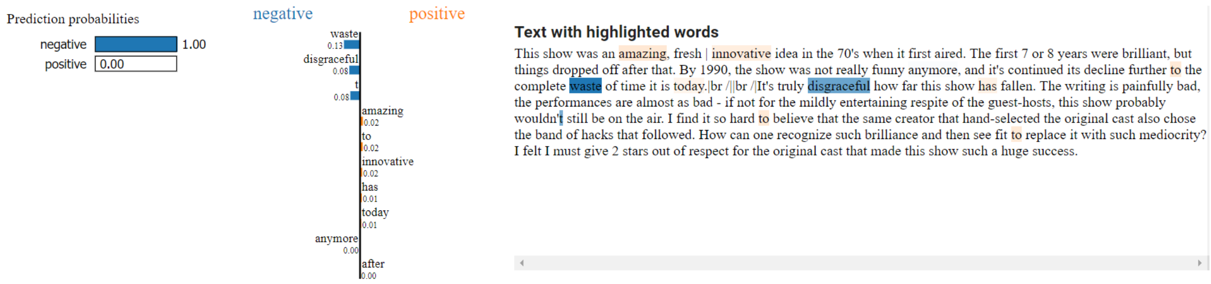
```
model 1:
```



Text with highlighted words

This show was an amazing, fresh | innovative idea in the 70's when it first aired. The first 7 or 8 years were brilliant, but things dropped off after that. By 1990, the show was not really funny anymore, and it's continued its decline further to the complete waste of time it is today.|br /||br /|It's truly disgraceful how far this show has fallen. The writing is painfully bad, the performances are almost as bad - if not for the mildly entertaining respite of the guest-hosts, this show probably wouldn't still be on the air. I find it so hard to believe that the same creator that hand-selected the original cast also chose the band of hacks that followed. How can one recognize such brilliance and then see fit to replace it with such mediocrity? I felt I must give 2 stars out of respect for the original cast that made this show such a huge success.

```
model 2:
```

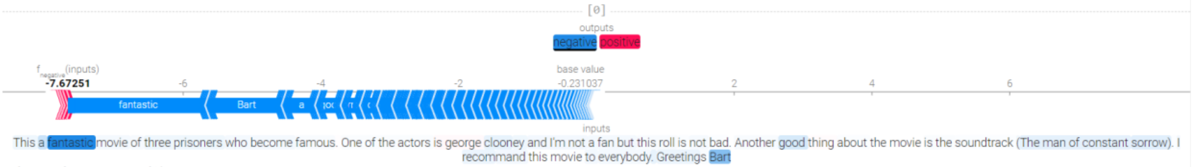


Text with highlighted words

This show was an amazing, fresh | innovative idea in the 70's when it first aired. The first 7 or 8 years were brilliant, but things dropped off after that. By 1990, the show was not really funny anymore, and it's continued its decline further to the complete waste of time it is today.|br /||br /|It's truly disgraceful how far this show has fallen. The writing is painfully bad, the performances are almost as bad - if not for the mildly entertaining respite of the guest-hosts, this show probably wouldn't still be on the air. I find it so hard to believe that the same creator that hand-selected the original cast also chose the band of hacks that followed. How can one recognize such brilliance and then see fit to replace it with such mediocrity? I felt I must give 2 stars out of respect for the original cast that made this show such a huge success.

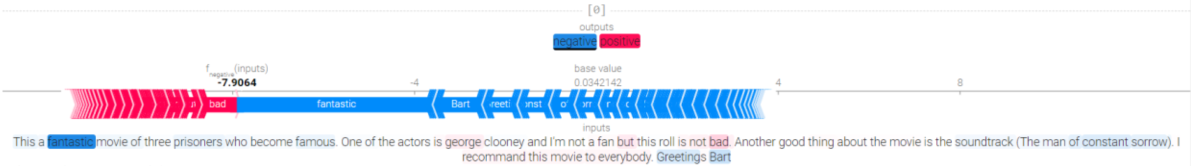
- **SHAP**

Long 1



This a fantastic movie of three prisoners who become famous. One of the actors is george clooney and I'm not a fan but this roll is not bad. Another good thing about the movie is the soundtrack (The man of constant sorrow). I recommand this movie to everybody. Greetings Bart

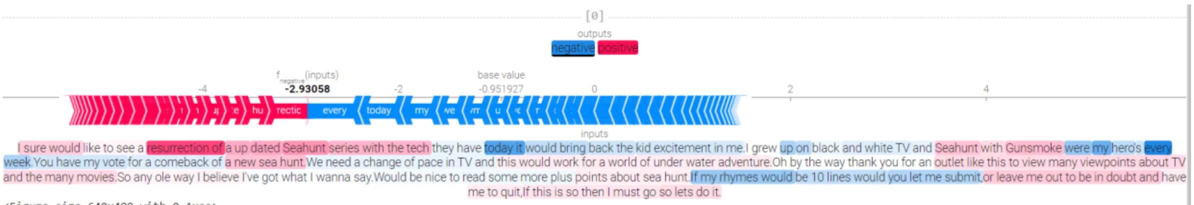
```
<Figure size 640x480 with 0 Axes>
model 2:
```



This a fantastic movie of three prisoners who become famous. One of the actors is george clooney and I'm not a fan but this roll is not bad. Another good thing about the movie is the soundtrack (The man of constant sorrow). I recommand this movie to everybody. Greetings Bart

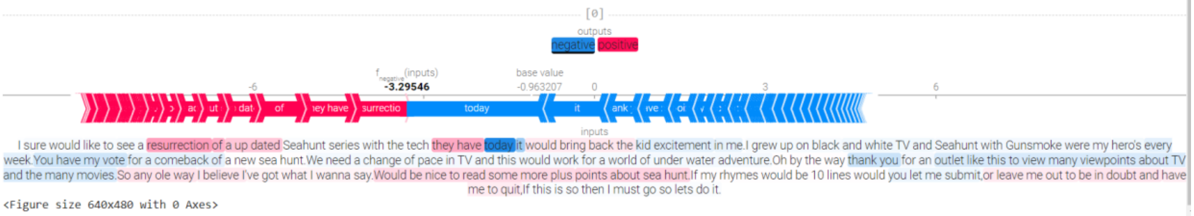
```
<Figure size 640x480 with 0 Axes>
```

Long 2



I sure would like to see a resurrection of a up dated Seahunt series with the tech they have today it would bring back the kid excitement in me.I grew up on black and white TV and Seahunt with Gunsmoke were my hero's every week.You have my vote for a comeback of a new sea hunt.We need a change of pace in TV and this would work for a world of under water adventure.Oh by the way thank you for an outlet like this to view many viewpoints about TV and the many movies.So any ole way I believe I've got what I wanna say.Would be nice to read some more plus points about sea hunt.If my rhymes would be 10 lines would you let me submit,or leave me out to be in doubt and have me to quit,If this is so then I must go so lets do it.

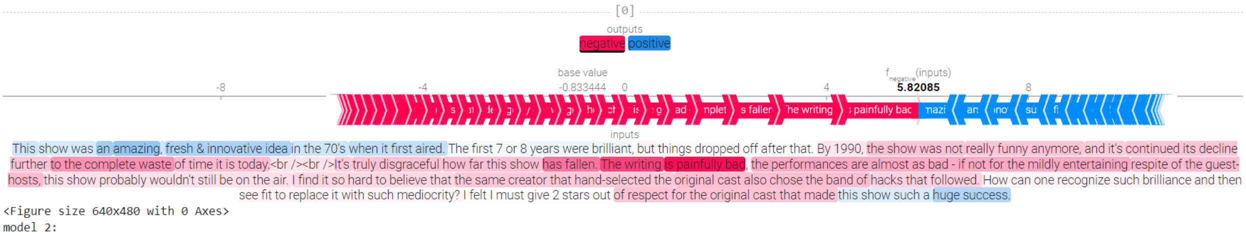
```
<Figure size 640x480 with 0 Axes>
model 2:
```



I sure would like to see a resurrection of a up dated Seahunt series with the tech they have today it would bring back the kid excitement in me.I grew up on black and white TV and Seahunt with Gunsmoke were my hero's every week.You have my vote for a comeback of a new sea hunt.We need a change of pace in TV and this would work for a world of under water adventure.Oh by the way thank you for an outlet like this to view many viewpoints about TV and the many movies.So any ole way I believe I've got what I wanna say.Would be nice to read some more plus points about sea hunt.If my rhymes would be 10 lines would you let me submit,or leave me out to be in doubt and have me to quit,If this is so then I must go so lets do it.

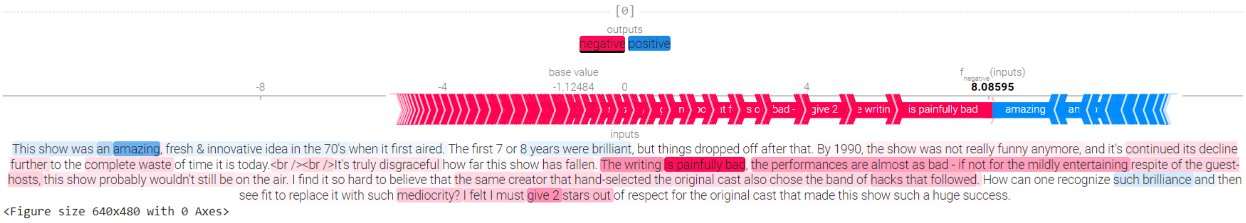
```
<Figure size 640x480 with 0 Axes>
```

Long 3



This show was an amazing, fresh & innovative idea in the 70's when it first aired. The first 7 or 8 years were brilliant, but things dropped off after that. By 1990, the show was not really funny anymore, and it's continued its decline further to the complete waste of time it is today.<br /><br />It's truly disgraceful how far this show has fallen. The writing is painfully bad, the performances are almost as bad - if not for the mildly entertaining respite of the guest-hosts, this show probably wouldn't still be on the air. I find it so hard to believe that the same creator that hand-selected the original cast also chose the band of hacks that followed. How can one recognize such brilliance and then see fit to replace it with such mediocrity? I felt I must give 2 stars out of respect for the original cast that made this show such a huge success.

```
<Figure size 640x480 with 0 Axes>
model 2:
```



This show was an amazing, fresh & innovative idea in the 70's when it first aired. The first 7 or 8 years were brilliant, but things dropped off after that. By 1990, the show was not really funny anymore, and it's continued its decline further to the complete waste of time it is today.<br /><br />It's truly disgraceful how far this show has fallen. The writing is painfully bad, the performances are almost as bad - if not for the mildly entertaining respite of the guest-hosts, this show probably wouldn't still be on the air. I find it so hard to believe that the same creator that hand-selected the original cast also chose the band of hacks that followed. How can one recognize such brilliance and then see fit to replace it with such mediocrity? I felt I must give 2 stars out of respect for the original cast that made this show such a huge success.

```
<Figure size 640x480 with 0 Axes>
```

- **綜合比較**

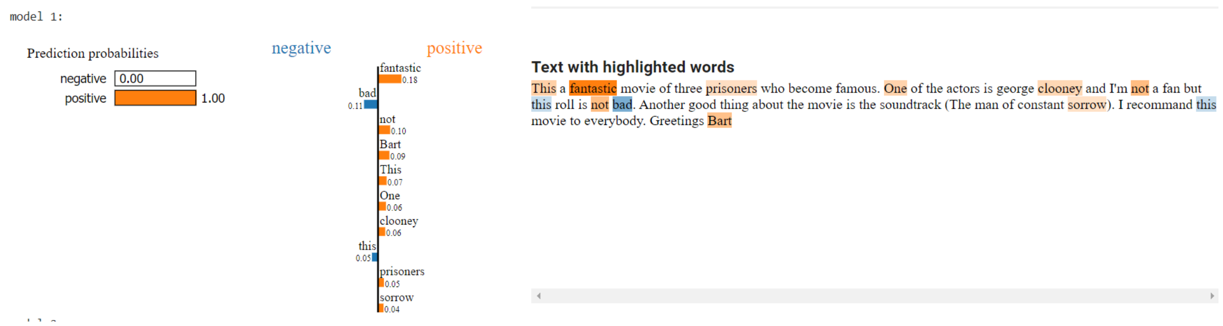
  - **Model 1 & 2**

    由上面的 model 1 & 2 的結果比較下來。我覺得 2 個 model 的表現似乎都差不多。都表現得不錯。但就 LIME 的 long 2 中可以看到 model 2 預測 positive 的機率高於 model 1。因此我認為也許 model 2 相較於 model 1 有好一點。

  - **LIME & SHAP**

根據上述結果，可以觀察雖然 LIME 和 SHAP 都能夠很好地解釋短句，但是在處理 IMDB 中長句的解釋時，LIME 的解釋效果較差，而 SHAP 的解釋效果較好。LIME 只能標記一些單詞，而無法提供整個句子的解釋，而且有些 LIME 標記的單詞有的有奇怪。像是”long 2”的句子中，LIME 標記了”would”為負面用詞，也會標記了影集的名字” Seahunt” 為負面用詞。相較之下，SHAP 的解釋效果就要好得多。SHAP 會逐詞標記單詞，並且那些具有較高 Shapley 值的單詞確實是評論中否定性的主要原因。此外，SHAP 還能夠標記負評中的正面部分。因此，在情感分類方面，我認為 SHAP 是比 LIME 更好的解釋方法。

我認為 LIME 對於長句解釋較差的原因主要在於，LIME 將評論中的每個詞都視為一個特徵。當特徵數量太多時，生成穩定的擾動數據集變得比較困難。因此若是特徵太多，則很難生成穩定的擾動數據集。而且，我們無法訓練一個簡單的分類器來近似原始模型。另外 LIME 的另一個缺點是他的擾動數據集是隨機的，這可能導致 LIME 給出的解釋不一致。如下圖所示，對於”Long 1” 的標註特徵與前面”Long 1” 的標註特徵不同。



```
model 1:
```

Text with highlighted words

This a fantastic movie of three prisoners who become famous. One of the actors is george clooney and I'm not a fan but this roll is not bad. Another good thing about the movie is the soundtrack (The man of constant sorrow). I recommand this movie to everybody. Greetings Bart

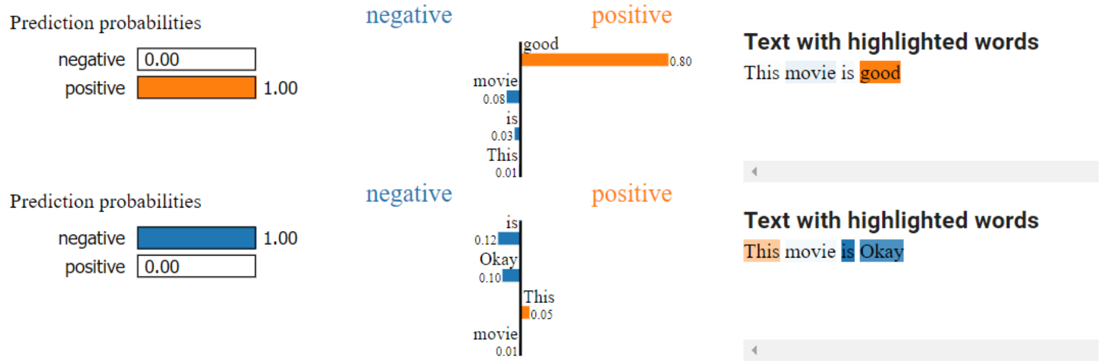
3. Try 3 different input sentences for attacks. Also, describe your findings and how to prevent the attack if you retrain the model in the future.

在這個部份，我會嘗試攻擊 TA 給出的模型(TA_model_1.pt)。

3.1 替換同義字

Text: This movie is good → This movie is okay

Text: This movie is about a dog → This movie is about a cat



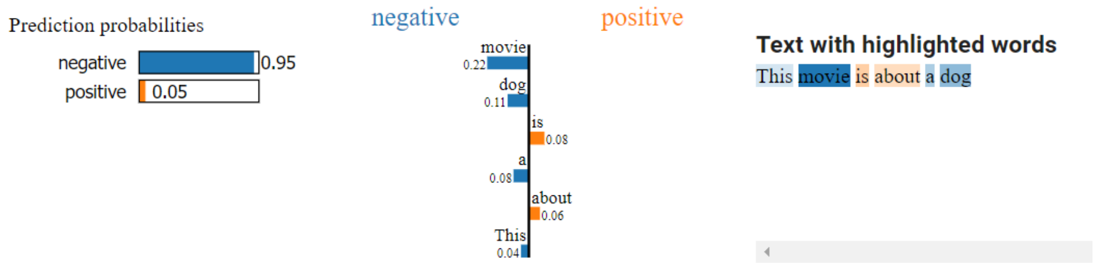
Text with highlighted words

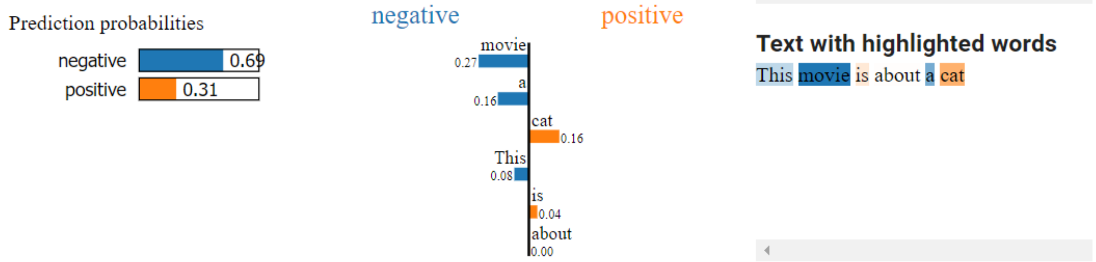
This movie is good

Text with highlighted words

This movie is Okay

這個句子更改為情緒較低落的同義詞。但正如上圖中看到，模型預測它是負評的。也許是因為助教的 model 在訓練時沒有使用沒有明顯情緒的句子，或者它的模型本身覺得“okay”是一個負面詞。

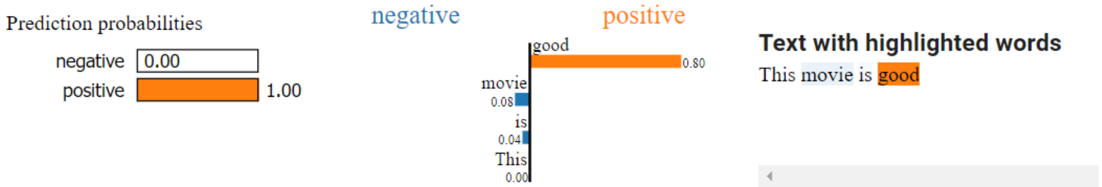


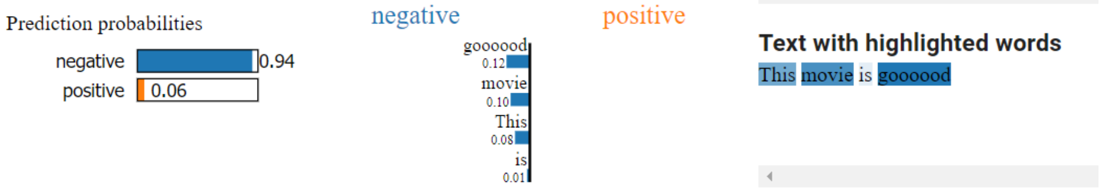


由圖中可看出，我只有把 dog 換成 cat，雖然模型的預測結果仍是一樣，但是可以從預測機率的狀況看出，模型把 dog 當成了負面詞，把 cat 當成正面詞。雖然還不構成很有效的攻擊，但是有用。

3.2 錯字

Text: This movie is good → This movie is goooood

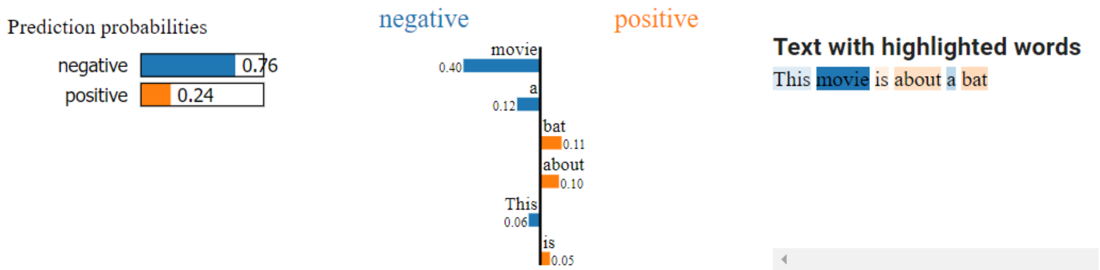




由上圖可以看出將“good”改拼成“goooood”。雖然我只在 good 中多加了暨個“o”。但 BERT 的輸出從 100% positive 變為 94% negative。

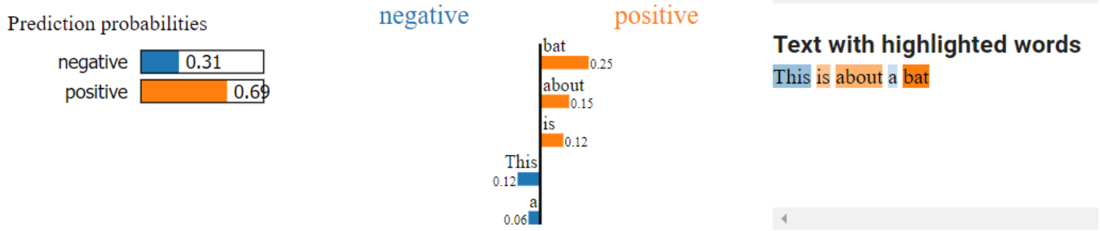
3.3 刪字

Text: This movie is about a bat → This is about a bat



Text with highlighted words

This movie is about a bat



Text with highlighted words

This is about a bat

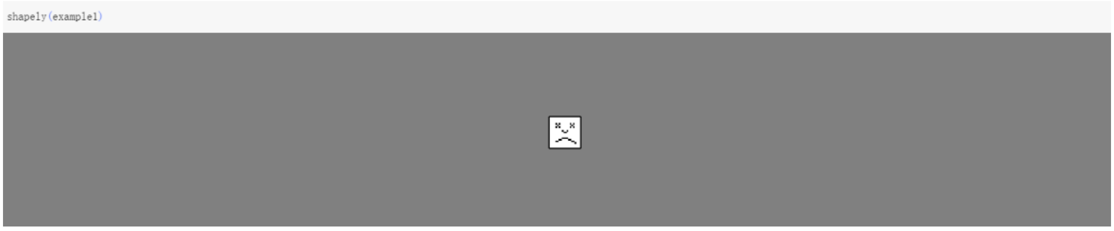
由上圖可以觀察出，把 movie 去除掉之後，模型的預測結果也大改變，由正面變為負面，我也不太懂為何模型會把 movie 當成是負面詞。

由上述實驗下來我認為要預防 NLP 攻擊的方式，應該可以嘗試在重新訓練 NLP 模型時，增加更多具有多樣性的攻擊樣本，包括同義詞替換、詞語刪除和字符層面的轉換。同時也要提高模型對於否定詞、情感變化和語義關係的理解。或是定期更新訓練數據，來加強模型的應變能力和防禦能力。

Bonus

- **Describe problems you meet and how you solve them.**

1. 有的時候不知道為什麼我的 colab 在要跑結果圖時，他說已經跑完了，但圖片會跑不出來或是呈現下面死機的狀況@@。都是靠重開才解決。



```
shapely(example1)
```

2. IMDB 的評論有的會同時含有’以及” ，很煩。原本都打算一個一個用跳脫符號去掉，但之後覺得這樣好白癡。所有就索性只找只使用’的評論(或是都不用)。之後就用”把他們刮起來就好!