AI Homework 4

109550127 宋哲頤

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

在ExBert中我特別覺得有兩個功能很厲害，第一個是自由選擇layer，通過不同的layer來觀察Model的變化，從distilbert-base-uncased的1~6 layers可以看到每個字被處理過後的變化。第二個是他可以每個字被訓練時的dataset是什麼，哪些Data影響了某個字的預測和磁性，而通過切換每一層查看該字在背後的dataset是怎麼被處理的。另外，當選擇特定的某單字時，可以看到其他字有多少head，這說明了影響的比例。

2. Compare at least 2 sentiment classification models (e.g.,TA_model_1.pt, your model in HW2).

預設example1為positive、example為negative

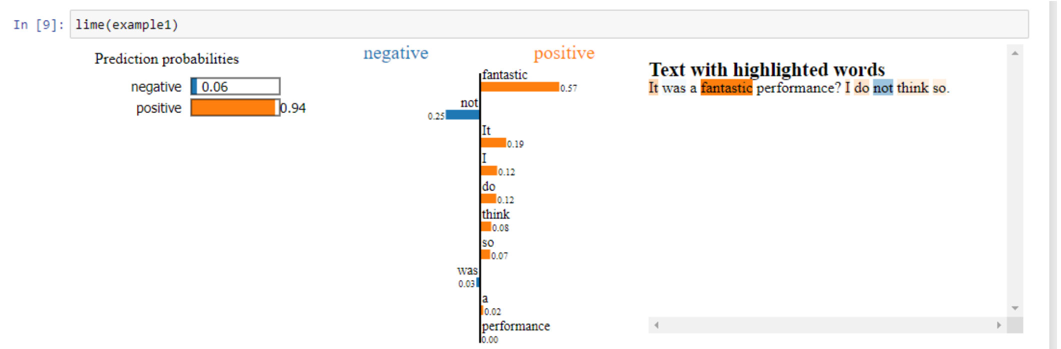
(1)

example1 = 'It was a fantastic performance? I do not think so.'

example2 = 'That is a fantastic movie? It is the most terrible movie that I have ever seen .'

跑完兩個model後可見倆著判斷的結果是差不多的，因fantastic一詞讓他

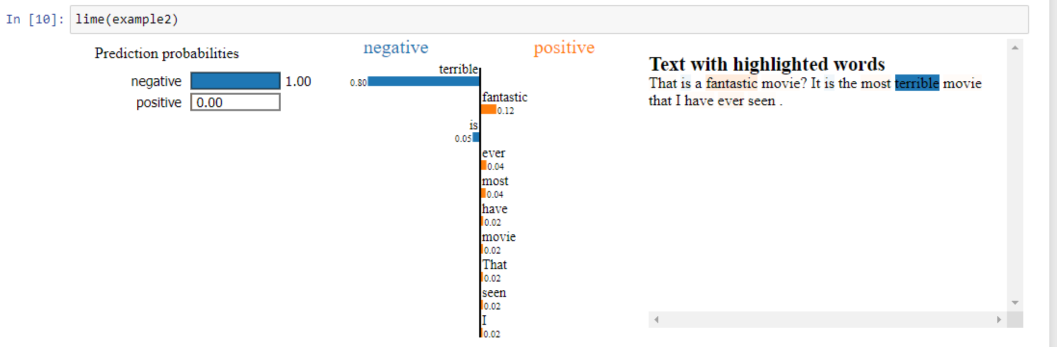
們認為是positive由not產生了一點negative，但其實整句話翻譯應該是中間偏negative，從model2可見對not的影響比較大，而example2的terrible影響力太大了，儘管都有fantastic但都鑾是大大的預測是negative，應該是dataset裡的terrible超高機率都是negative的緣故。



```
In [9]: lime(example1)
```

Text with highlighted words

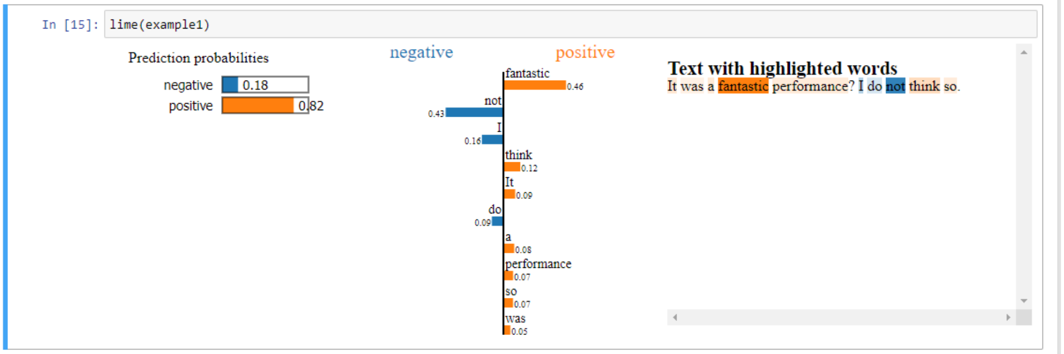
It was a fantastic performance? I do not think so.



```
In [10]: lime(example2)
```

Text with highlighted words

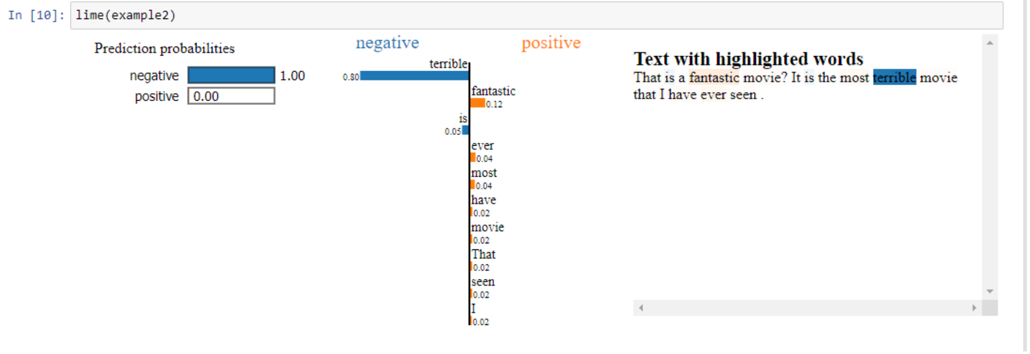
That is a fantastic movie? It is the most terrible movie that I have ever seen .



```
In [15]: lime(example1)
```

Text with highlighted words

It was a fantastic performance? I do not think so.



```
In [10]: lime(example2)
```

Text with highlighted words

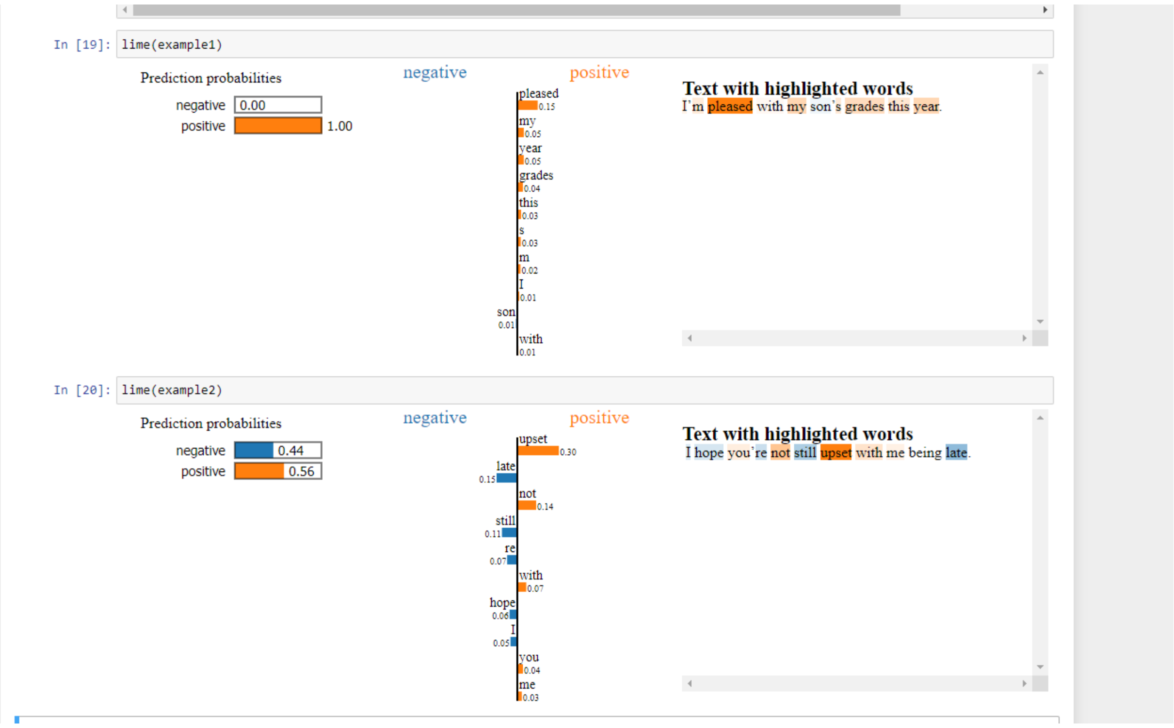
That is a fantastic movie? It is the most terrible movie that I have ever seen .

(2)

example1 = 'I’ m pleased with my son’ s grades this year.'

example2 = ' I hope you’ re not still upset with me being late.'

這兩個example不是電影評論的句子可以完全看出來model是沒有辦法分辨好或壞，最令我驚訝的是兩個model的upset都是positive，一般評論講到upset居然都是正面的?另外model2的son跟grade都是negative而model是positive也是很有趣。或許model2講到grade都認為是對電影負評的分數吧。



```
In [19]: lime(example1)
```

Prediction probabilities

negative	0.00
positive	1.00

Text with highlighted words

I’m pleased with my son’s grades this year.

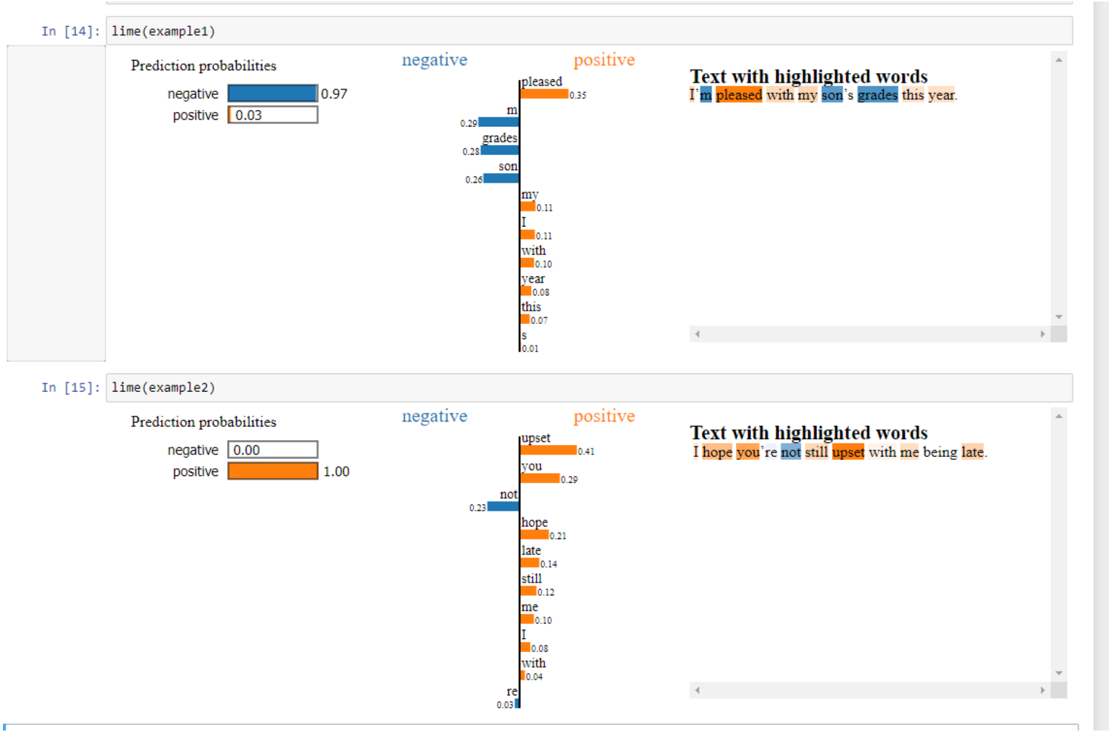
```
In [20]: lime(example2)
```

Prediction probabilities

negative	0.44
positive	0.56

Text with highlighted words

I hope you’re not still upset with me being late.



```
In [14]: lime(example1)
```

Text with highlighted words

I'm pleased with my son's grades this year.

```
In [15]: lime(example2)
```

Text with highlighted words

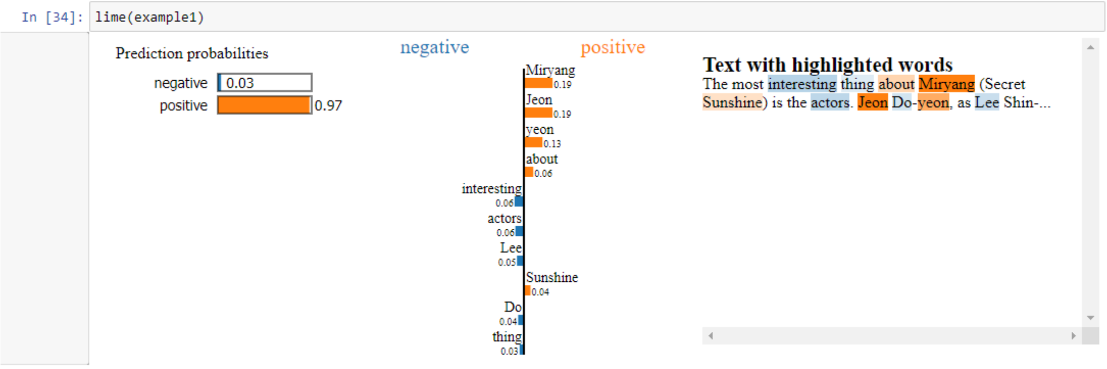
I hope you're not still upset with me being late.

(3)

example1 = 'The most interesting thing about Miryang (Secret Sunshine) is the actors. Jeon Do-yeon, as Lee Shin-...'

example2 = 'This movie was so poorly written and directed I fell asleep 30 minutes through the movie. The jokes ...'

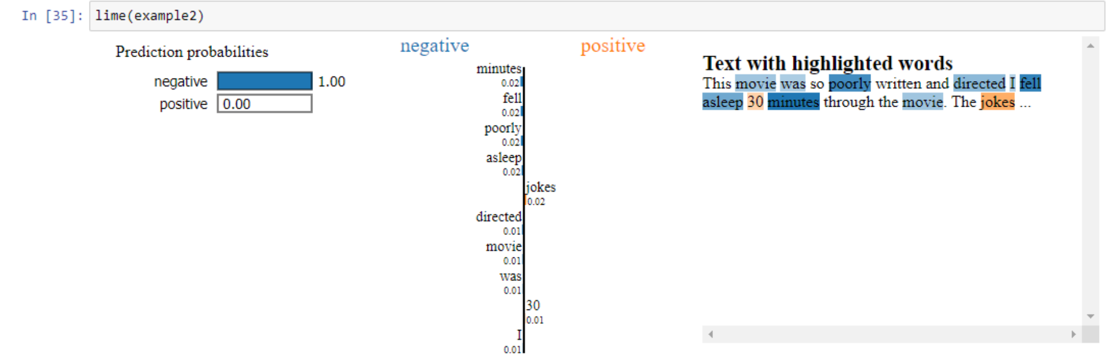
這兩個例子我是從網路上的其他電影評論dataset中選的，我嘗試了不將完整內容完全複製來看看結果，在成果上兩個model都表現得很正確，而例如一些地方名詞或人名都會判斷為positive，interesting之類的看似正面形容詞則判斷為負，整體上兩個model相差無幾，判斷這次例子中的關鍵字的方式也很像。



```
In [34]: lime(example1)
```

Text with highlighted words

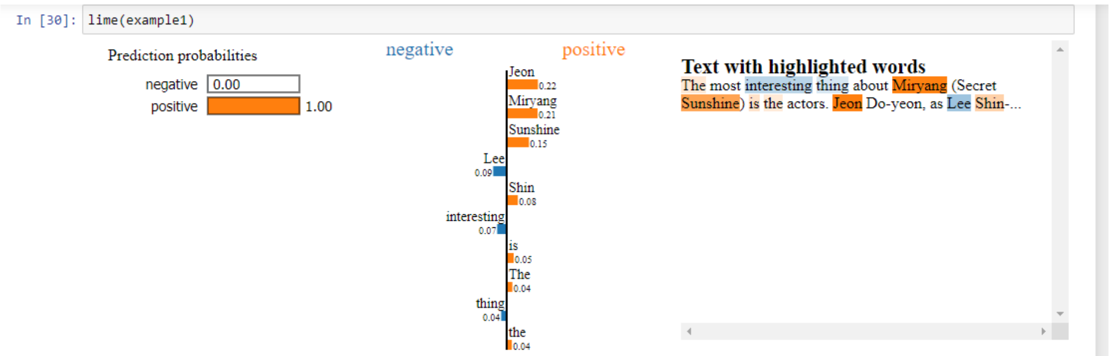
The most interesting thing about Miryang (Secret Sunshine) is the actors. Jeon Do-yeon, as Lee Shin-...



```
In [35]: lime(example2)
```

Text with highlighted words

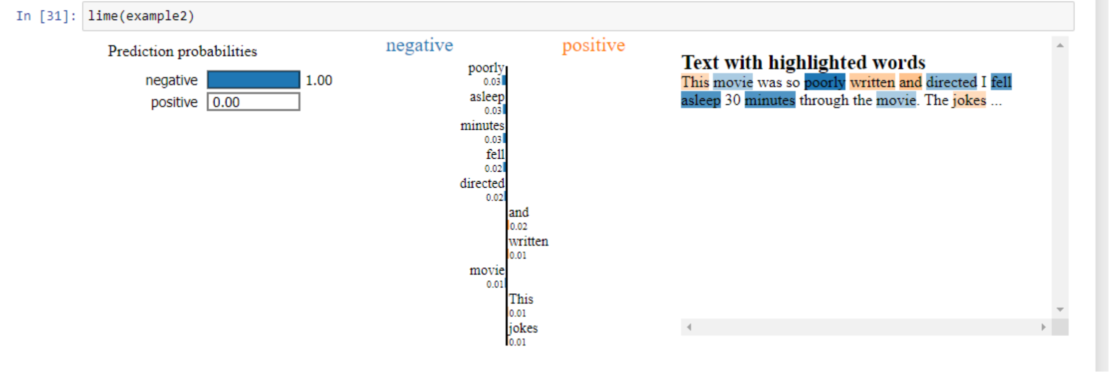
This movie was so poorly written and directed I fell asleep 30 minutes through the movie. The jokes ...



```
In [30]: lime(example1)
```

Text with highlighted words

The most interesting thing about Miryang (Secret Sunshine) is the actors. Jeon Do-yeon, as Lee Shin-...



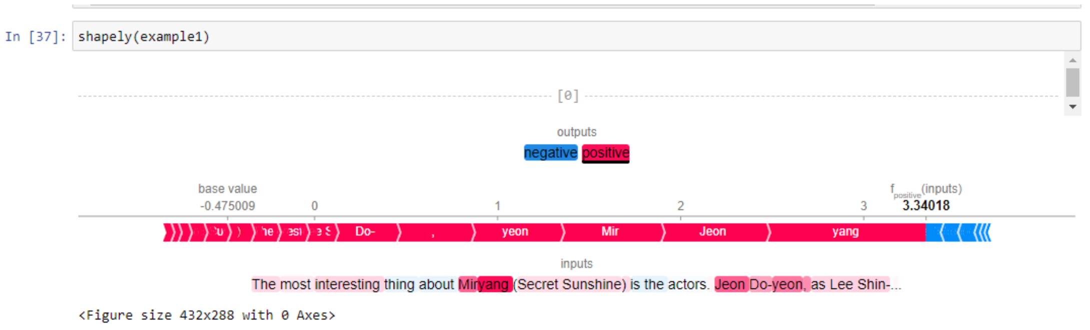
```
In [31]: lime(example2)
```

Text with highlighted words

This movie was so poorly written and directed I fell asleep 30 minutes through the movie. The jokes ...

2. Compare the explanation of **LIME and SHAP**.

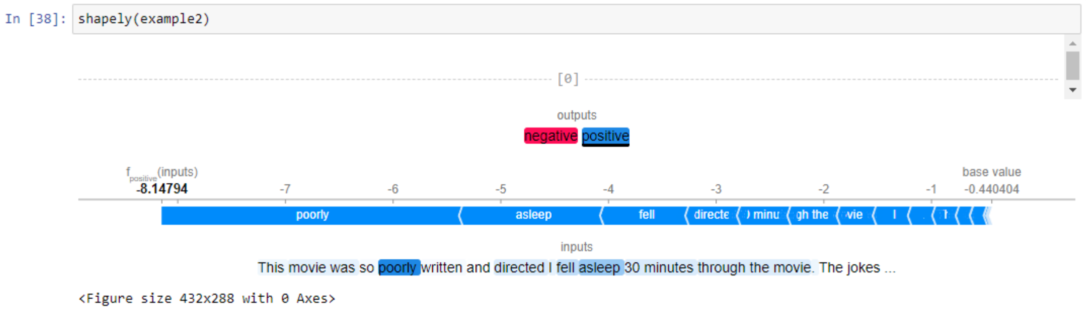
使用2.3的example和model1來看SHAP的explanation



In [37]:
```
shapely(example1)
```

[0]
outputs
negative positive
base value
-0.475009
0
1
2
3
f positive (inputs)
3.34018
Do-
yeon
Mir
Jeon
yang
inputs
The most interesting thing about Miryang (Secret Sunshine) is the actors. Jeon Do-yeon, as Lee Shin-...

```
<Figure size 432x288 with 0 Axes>
```



In [38]:
```
shapely(example2)
```

```
<Figure size 432x288 with 0 Axes>
```

LIME是通過建立一個局部代理模型來近似解釋黑盒模型的預測結果。且並不僅僅是基於模型的局部數據來構建代理模型，還會產生一些數據擾動 (data variations)，基於原始擾動的數據以及黑盒模型的預測值，建立一個可解釋的白盒模型，比如Lasso，決策樹等。缺點: LIME在解釋文本模型的時候，最大的不足是結果的不穩定性。因為不同的局部取樣帶來不同的局部擾動，最終的解釋結果會有波動。LIME相對計算速度會比SHAP要快。

SHAP是一個通用性模型可解釋性框架，Shapley regression values在計算特征貢獻的時候會在特征子集上重新訓練模型。對於特征i，首先產生所有包含i和去除i的特征集合，然後重新訓練並計算預測結果，以此計算特征i的貢獻的平均。缺點: SHAP value的計算會非常困難和耗時，因此在SHAP框架中有多

個獨特的計算方式，Kernel SHAP是一個通用解釋算法，Deep SHAP用於計算深度學習模型的SHAP value。

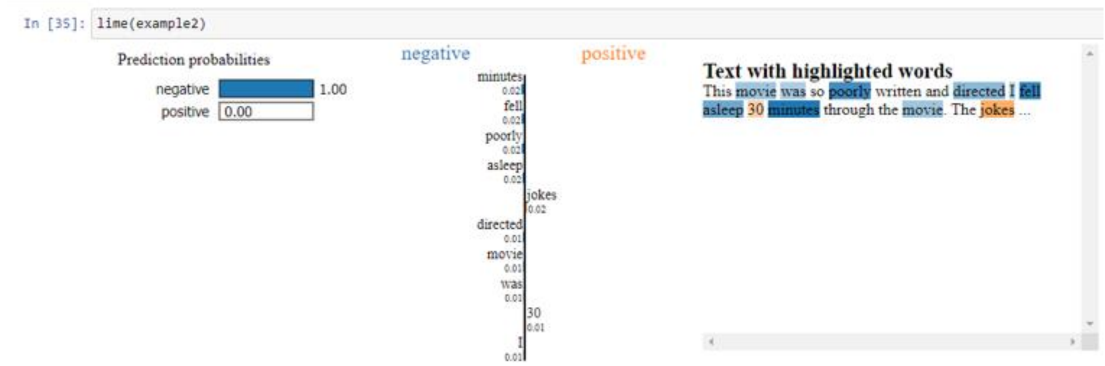
以上述的example來說，兩個解釋性模型的結果都差不多，可以看到的是SHAP比較偏向用一段小句子來判斷，而LIME偏向是一個字一個字判斷，沒有誰好誰壞的問題，以結果來說都是正確的。

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.

(1)改變詞性

example = 'This movie was so poorly written and directed I fell asleep 30 minutes through the movie. The jokes ...'

來說，我把movie加s、directed便direct、minutes少s

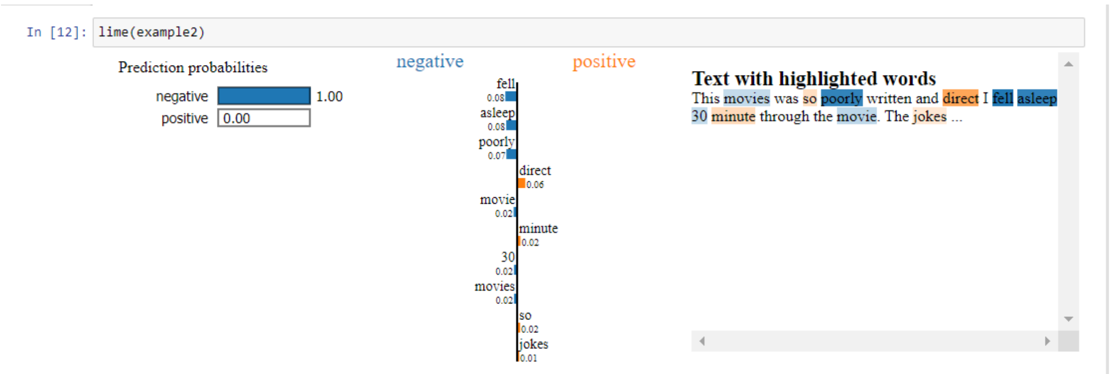


```
In [35]: lime(example2)
```

Text with highlighted words

This movie was so poorly written and directed I fell asleep 30 minutes through the movie. The jokes ...

example: This movies was so poorly written and direct I fell asleep 30 minute through the movie. The jokes ...



```
In [12]: lime(example2)
```

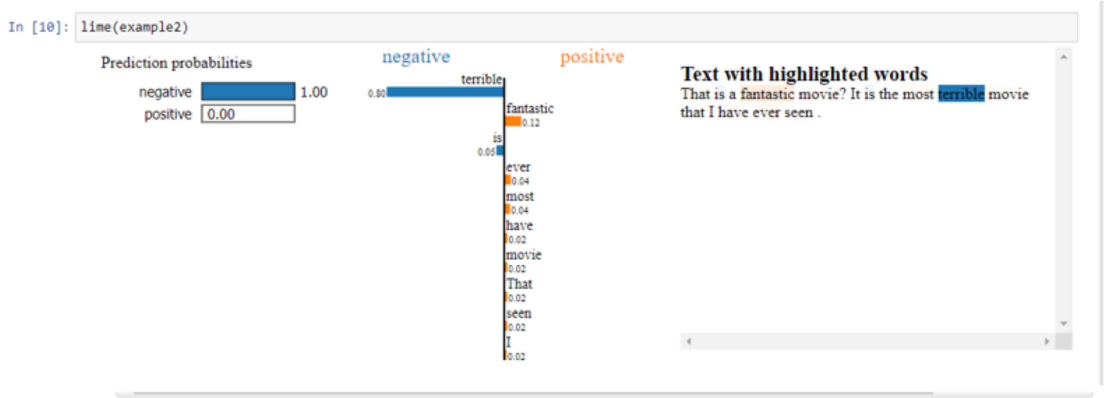
Text with highlighted words

This movies was so poorly written and direct I fell asleep 30 minute through the movie. The jokes ...

可見direct跟minute的預測都相反了，僅僅改變詞性跟單複數就對explain有attack。

(2)swap

example= 'That is a fantastic movie? It is the most terrible movie that I have ever seen .'



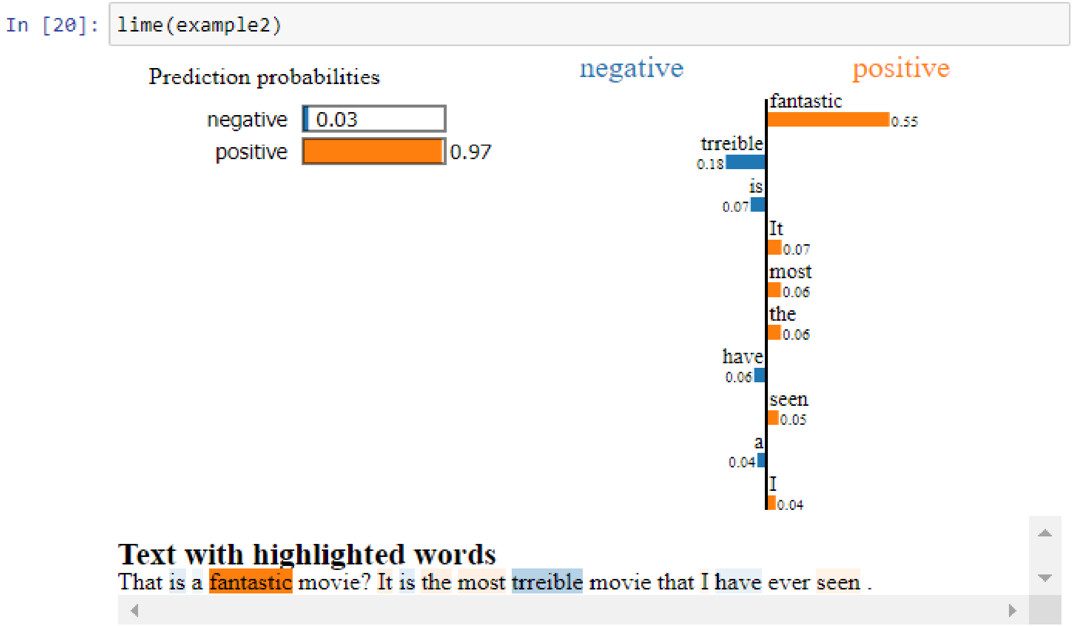
```
In [10]: lime(example2)
```

Text with highlighted words

That is a fantastic movie? It is the most terrible movie that I have ever seen .

example= 'That is a fantastic movie? It is the most trreible movie that I have ever seen .'

同樣的句子我將terrible變成trreible後，他的影響力小了許多，甚至原本預測negative變成positive，讓fantastic的影響力改變它的結果。



```
In [20]: lime(example2)
```

Prediction probabilities

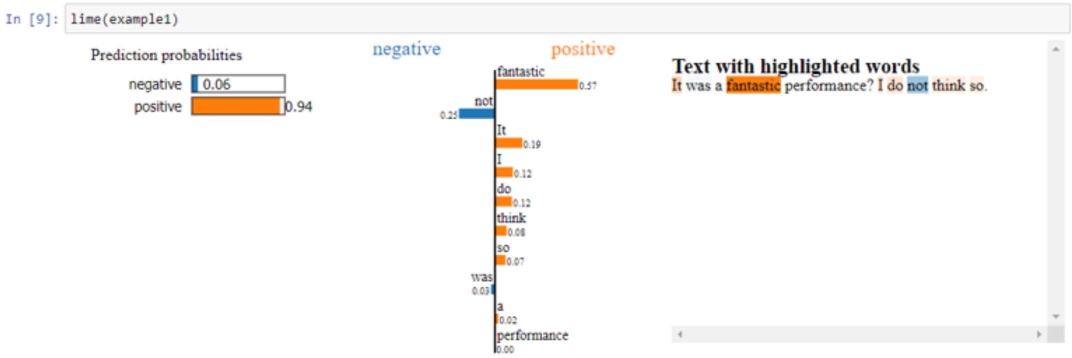
negative 0.03
positive 0.97

Text with highlighted words

That is a fantastic movie? It is the most trreible movie that I have ever seen .

(3)delete

example = 'It was a fantastic performance? I do not think so.'



```
In [9]: lime(example1)
```

Prediction probabilities

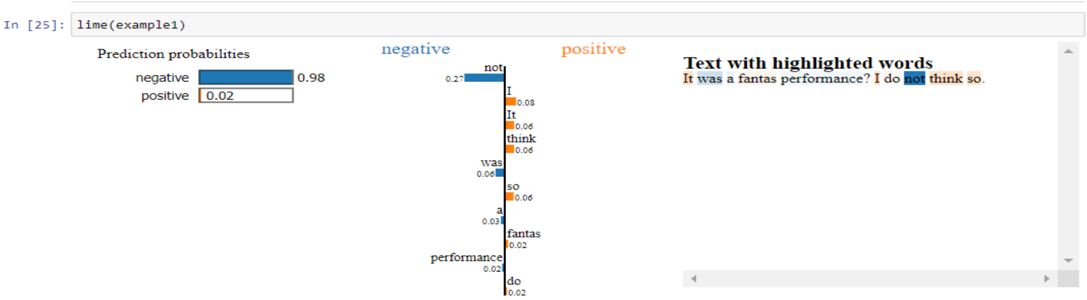
negative 0.06
positive 0.94

Text with highlighted words

It was a fantastic performance? I do not think so.

example= 'It was a fantas performance? I do not think so.'

將fantastic刪除一些字變fantas，而LIME並沒有將fantas和fantastic聯繫上，造成影響力小了許多，讓其他字把預測從positive變negative。



```
In [25]: lime(example1)
```

Prediction probabilities

negative 0.98
positive 0.02

Text with highlighted words

It was a fantas performance? I do not think so.

(4)prevent the attack if you retrain the model in the future.

我想可以在model裡先把不合文法的句子或單字先去除掉不列入判斷，這樣可以減少詞性改變的attack。另外，如果有單字或句子有缺損也要篩掉，畢竟根本不存在的句子或單字對model的train根本沒意義。