AI hw4 Report

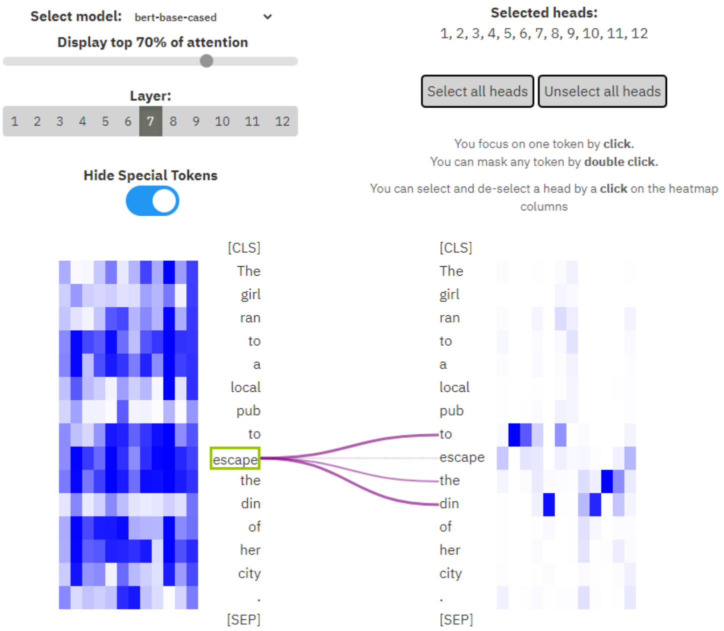
exBERT

1.1 Attention Mechanism of BERT

BERT 是傳統語言模型的一種變形，而語言模型做的事情就是在給定一些詞彙的前提下，去估計下一個詞彙出現的機率分佈，BERT 的訓練方法為預先訓練 (Pretraining) 和微調 (Fine-tune)。

預先訓練要求只利用序列本身已有的資訊即可，不需要進行人工標註，訓練任務有兩個，一是克漏字，隨機將語句中的詞語遮住並訓練 BERT 猜出該詞語為何，二是預測下一句，將其中一個句子以固定比率替換成其他句子，訓練 BERT 判斷該句子是否為上一個句子的下一句。

然而我們並不常使用克漏字和預測下一個句子，文本分類、摘要這些反而更常被使用，這時我們便需要做一些微調，只要在 BERT 輸出的部分訓練一個自己的小模型，整個模型的輸出便可符合需求。



Select model: bert-base-cased

Display top 70% of attention

Layer: 7

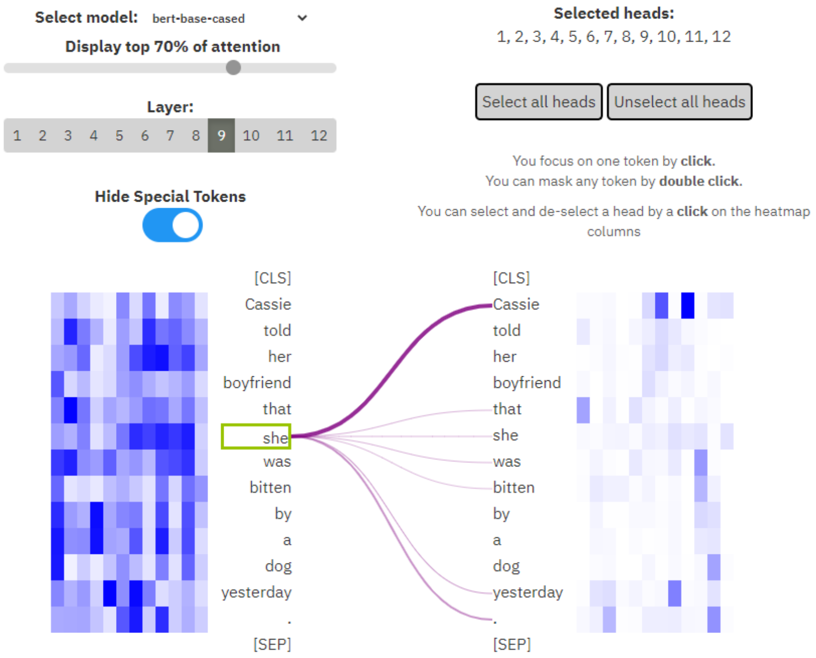
Selected heads: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

You focus on one token by **click.**
You can mask any token by **double click.**

You can select and de-select a head by a **click** on the heatmap columns

Language Modal：bert-base-cased
Input Sentence：The girl ran to a local pub to escape the din of her city.
MASK：escape

如上圖所示，如果我們將 escape 遮住，讓 BERT 去預測哪個詞應該出現在這裡，在 BERT 第五層中，最引起注意的是“to”,”the”,”din”這些和 escape 較為相關的詞語，這就是 attention mechanism 的例子之一。



Select model: bert-base-cased

Display top 70% of attention

Layer: 1 2 3 4 5 6 7 8 9 10 11 12

Hide Special Tokens

Selected heads:
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

Select all heads | Unselect all heads

You focus on one token by **click.**
You can mask any token by **double click.**

You can select and de-select a head by a **click** on the heatmap columns

Language Modal：bert-base-cased
Input Sentence：Cassie told her boyfriend that she was bitten by a dog yesterday.
MASK：she

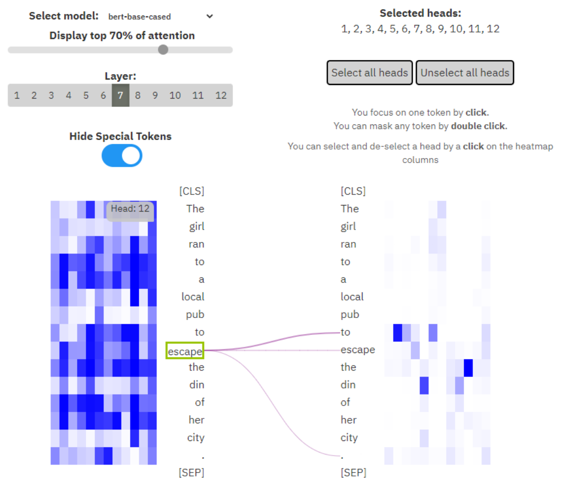
上圖說明了我將一個代名詞遮住，我們可以觀察到放置 ”Cassie”的機會是最高的，這說明了 BERT 能分析語法結構，並判斷出代名詞所指對象。

1.2 Comparison of BERT and DistilBERT

DistilBERT 是一個更小的 Transformer 模型，它與原始 BERT 模型有很多相似之處，DistilBERT 的參數大約只有 BERT 的 40%，運行速度快了 60%，還同時保留 BERT 99%的性能。

為了方便比較，我用相同的句子來測試，並遮住相同的詞語，句子如下：

“The girl ran to a local pub to escape the din of her city.”



Select model: bert-base-cased

Display top 70% of attention

Layer: 1 2 3 4 5 6 7 8 9 10 11 12

Hide Special Tokens

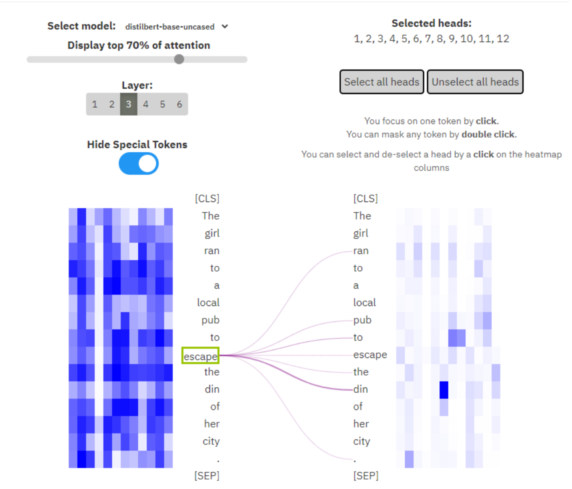
Selected heads:
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

Select all heads | Unselect all heads

You focus on one token by **click.**
You can mask any token by **double click.**

You can select and de-select a head by a **click** on the heatmap columns

bert-base-cased



Select model: distilbert-base-uncased

Display top 70% of attention

Layer: 1 2 3 4 5 6

Hide Special Tokens

Selected heads:
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

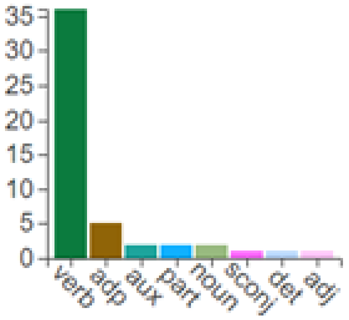
Select all heads | Unselect all heads

You focus on one token by **click.**
You can mask any token by **double click.**

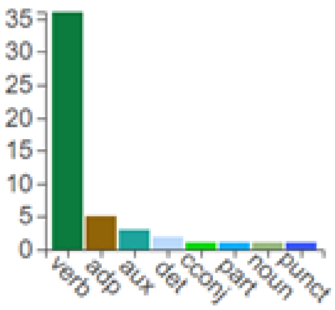
You can select and de-select a head by a **click** on the heatmap columns

distilbert-base-uncased

首先，我先將 bert-base-cased 和 distilbert-base-uncased 進行比較，可以發現 Distilbert 依然能夠準確預測，而且兩者關注的詞語也大同小異，這可以說明 Distilbert 確實是有些地方模仿原始的 BERT，並從那學習一些內容。



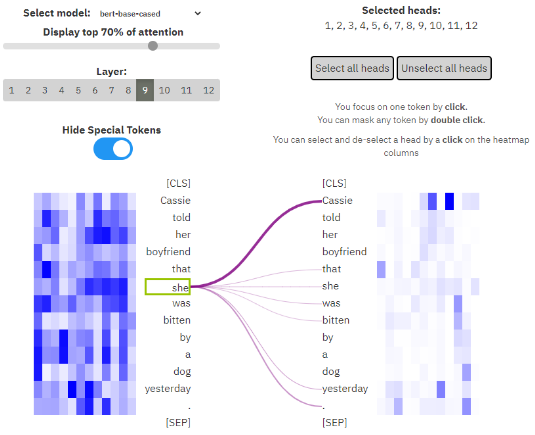
bert-base-cased



distilbert-base-uncased

exBERT 還能根據所選的詞語，去找出其他資料中的相同詞語，根據上圖我發現無論是 BERT 還是 DistilBERT，在此有著差不多的表現，escape 被判斷為動詞的機率大致上是相同的。

為了更精準的進行比較，我測試了另一個句子：
“Cassie told her boyfriend that she was bitten by a dog yesterday.”



Select model: bert-base-cased

Display top 70% of attention

Layer: 9

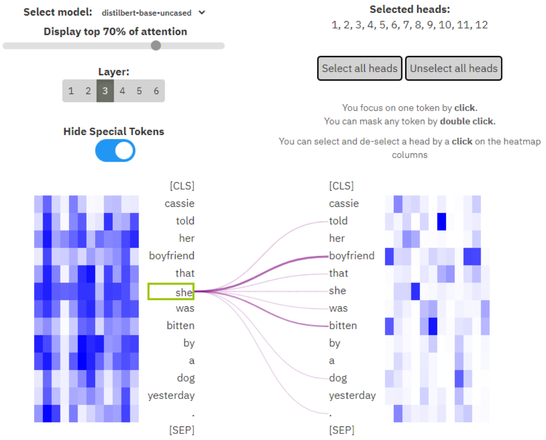
Hide Special Tokens

Selected heads: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

Select all heads | Unselect all heads

You focus on one token by click.
You can mask any token by double click.
You can select and de-select a head by a click on the heatmap columns

bert-base-cased



Select model: distilbert-base-uncased

Display top 70% of attention

Layer: 3

Hide Special Tokens

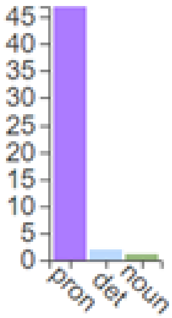
Selected heads: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

Select all heads | Unselect all heads

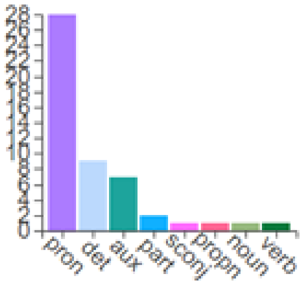
You focus on one token by click.
You can mask any token by double click.
You can select and de-select a head by a click on the heatmap columns

distilbert-base-uncased

根據上圖可以發現關於代名詞的判斷，雖然 Distilbert 有預測到一小部分的 she，但並沒有預測到 Cassie，BERT 則是兩者都有，這可以說明 BERT 似乎還是比 Distilbert 來的精準一些，至少 BERT 知道 she 指的是 Cassie，而不是 boyfriend (DistilBERT預測出來的結果)



bert-base-cased



distilbert-base-uncased

she 在日常生活中擔任的是代名詞的角色，根據上圖可以發現 she 在 BERT 的判斷中確實有九成是擔任代名詞的角色，但在 DistilBERT 大概只有一半多一點會被判斷是代名詞，這也說明了兩者的差異。

整體而言，如果是簡單的句子，我認為 DistilBERT 確實表現不錯，甚至是和 BERT 有幾乎相同的結果，但如果提升句子的難易度，讓句子結構複雜一些，DistilBERT 並不能有令人滿意的表現。

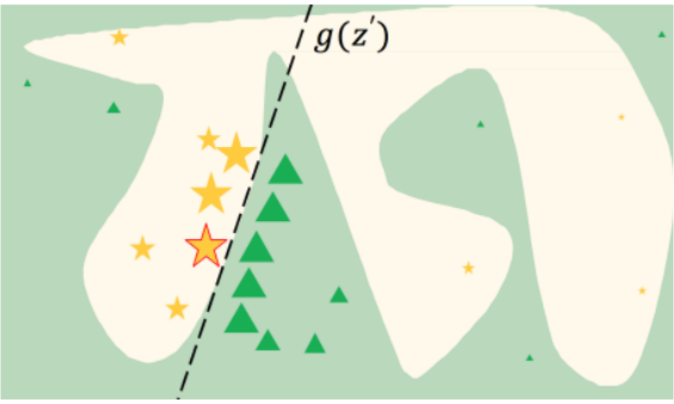
XAI：LIME and SHAP

2.1 Explainable AI

現在有很多AI模型已經被訓練得很好，但我們並不知道這些決策是如何被決定出來的，這會導致我們不相信AI所做的決策，除此之外，因為AI所做的決定不一定是正確的，缺乏可解釋性的模型如果做出了錯誤決策，可說是幾乎不可能調整的，所以我們需要 XAI 來讓使用者理解並信任機器學習演算法所建立的結果與輸出。

2.2 Local Interpretable Model-agnostic Explanations (LIME)

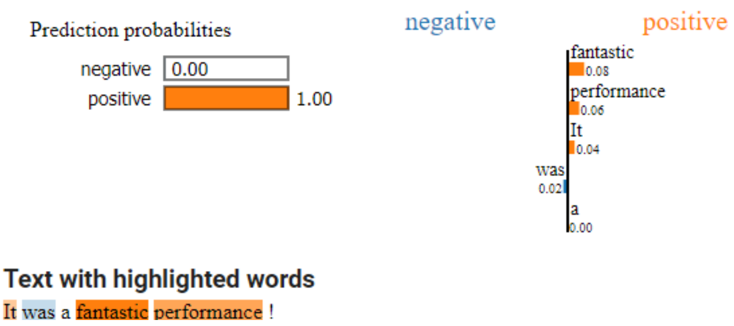
假設我們利用特徵 x_1 與 x_2 建立了一個複雜且準確的模型 **f (x)**去預測個體是位於黃色區域還是 綠色區域，從下圖可以發現，若我們想要解釋為什麼這個個體會被分類至黃色區域非常困難。



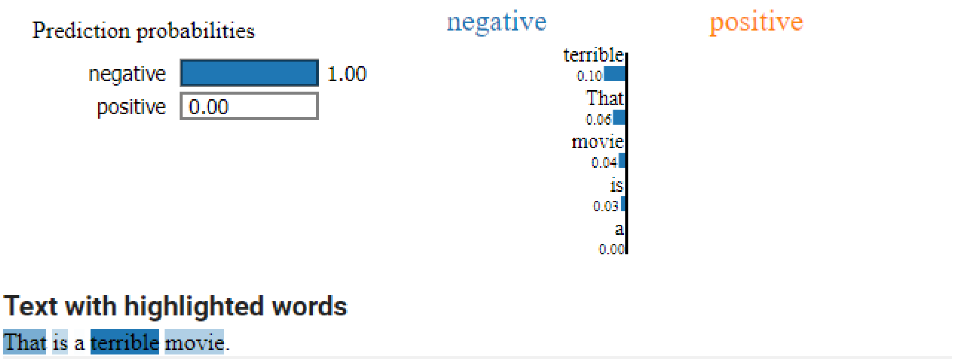
LIME 就是為了解決上述問題，在該個體的附近建立一個簡單可解釋的模型 ***g(z')*** (虛線)作為此個體的解釋模型，可解釋模型 ***g(z')*** 在該個體附近的預測準確度必須與原模型 ***f(x)*** 相當，但如果在距離該個體較遠的區域，其預測力就會大幅下降，這就是所謂的局部 (local) 忠實性。

LIME 的目標主要是針對一個複雜難解釋的模型，根據欲解釋的個體提供一個局部可解釋的模型，其主要的概念為針對某個體在局部區域找出一個簡單可理解的模型，用以回答「為什麼模型會將某個體分類到特定的類別」。

我使用助教所提供的模型 (TA_model_1.pt) 進行測試， 次模型為distilbert-base-uncased ，並用一些簡單的句子進行測試。



Sentence: It was a fantastic performance !



Sentence: That is a terrible movie.

根據上面兩張圖可以發現 LIME 可以準確解釋出主要是 fantastic 導致這個句子為正面的和主要是因為 terrible 進而判斷出句子為負面的，預測算是蠻準確的，但因為這是簡單的句子，所以不知道如果我將難度提升，LIME 是否還能準確地找出是那些字詞來決定句子是正面或是負面，我將在 2.4 將 LIME 和 SHAP 做一個更深入的比較。

2.3 SHapley Additive exPlanations (SHAP)

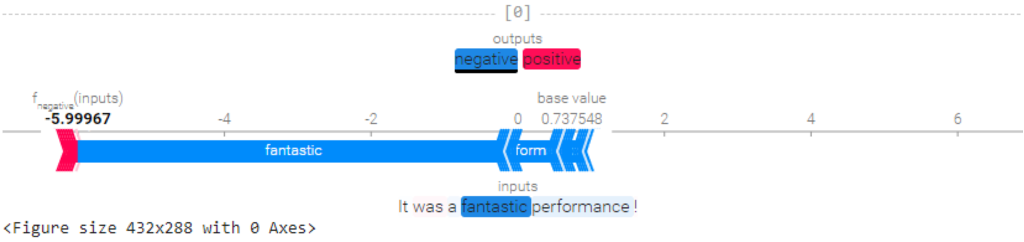
SHAP是一種基於Shapley Values 的方法，Shapley Values 是 Lloyd Stowell Shapley 基於合作賽局理論提出來解決方案，這種方法根據玩家們在遊戲中所得到的總支出，公平的分配總支出給玩家們，計算Shapley Values的公式如下:

$$\phi_j(val) = \sum_{S\subseteq\{x_1,..,x_p\}\setminus\{x_j\}} \frac{|S|!(p-|S|-1)!}{p!}\left(val(S\cup\{x_j\})-val(S)\right)$$

- x1, …, xp 為建立模型所使用的所有特徵，p 為所有的特徵數量
- S 為排除 xj 的子集合，可以看這個玩家的貢獻程度為何
- val(S) 是對於集合 S 的預測值減去期望值

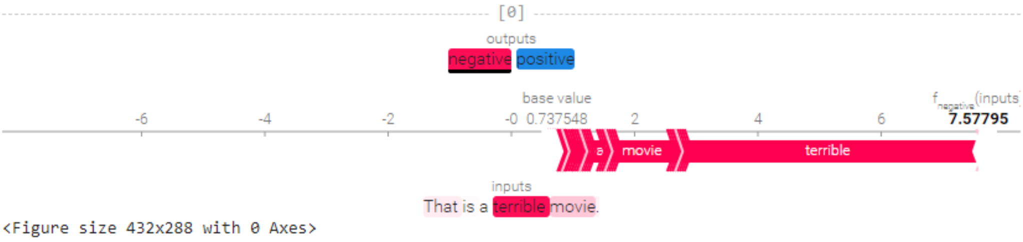
$$val_x(S) = \int \hat{f}(x_1, ..., x_p) d\mathbb{P}_{x \notin S} - E_x(\hat{f}(X))$$

在 SHAP，我依然使用 distilbert-base-uncased 和 2.2 的句子進行測試。



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

Sentence: It was a fantastic performance !



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

Sentence: That is a terrible movie.

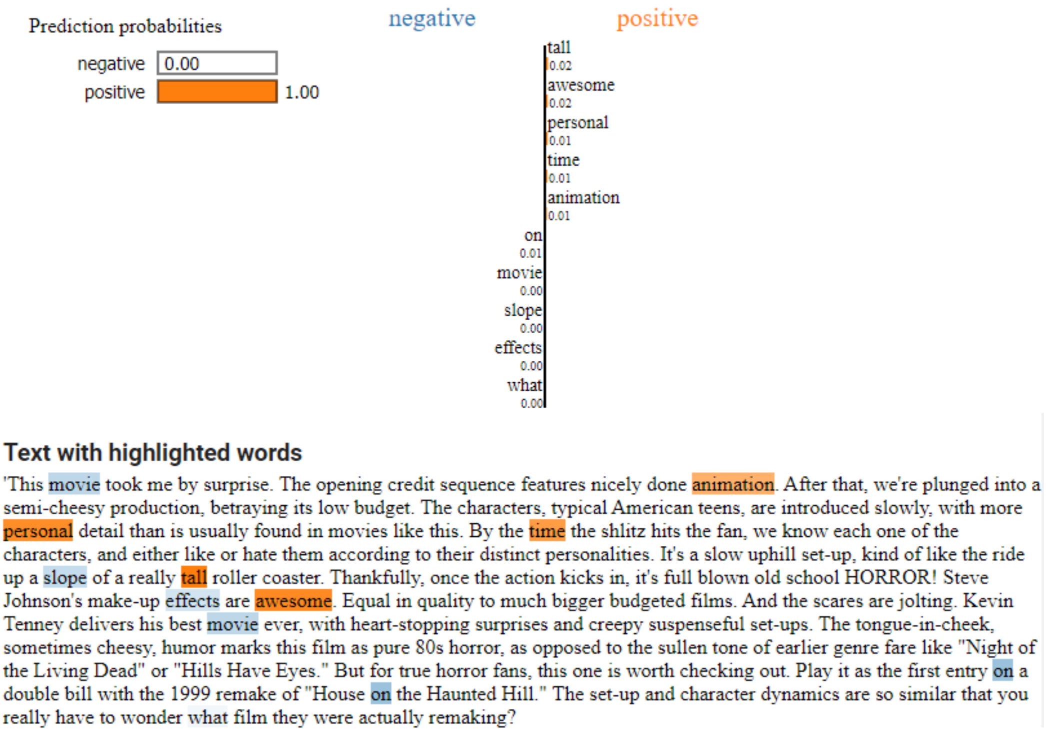
根據這兩張圖，我們發現 SHAP 和 LIME 一樣，都能正確解釋出主要是 fantastic 導致這個句子為正面的和主要是因為 terrible 進而判斷出句子為負面的。

2.4 Comparison of LIME and SHAP

為了更準確進行比較，我使用了四個我從 IMDB 中挑選出來的評論，分別為兩則正面評論和兩則負面評論，模型一樣是使用 distilbert-base-uncased。

1. This movie took me by surprise. The opening credit sequence features nicely done animation. After that, we're plunged into a semi-cheesy production, betraying its low budget. The characters, typical American teens, are introduced slowly, with more personal detail than is usually found in movies like this. By the time the shlitz hits the fan, we know each one of the characters, and either like or hate them according to their distinct personalities. It's a slow uphill set-up, kind of like the ride up a slope of a really tall roller coaster. Thankfully, once the action kicks in, it's full blown old school HORROR! Steve Johnson's make-up effects are awesome. Equal in quality to much bigger budgeted films. And the scares are jolting. Kevin Tenney delivers his best movie ever, with heart-stopping surprises and creepy suspenseful set-ups. The tongue-in-cheek, sometimes

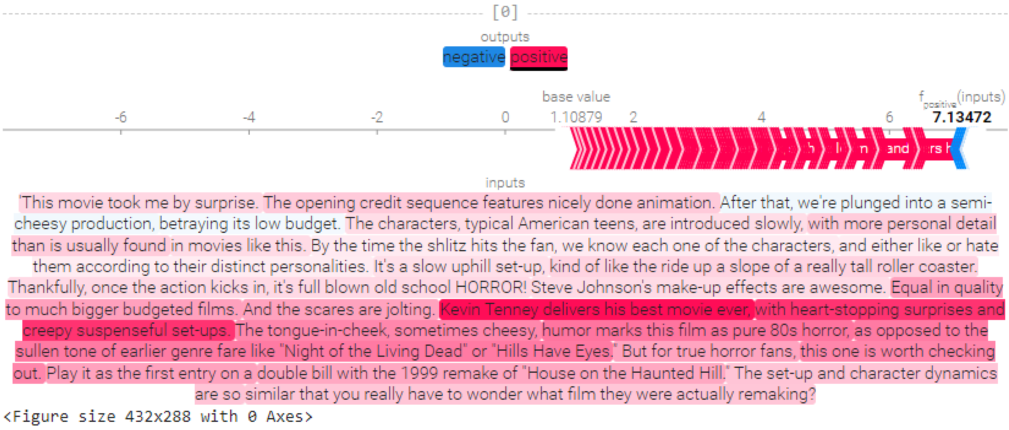
cheesy, humor marks this film as pure 80s horror, as opposed to the sullen tone of earlier genre fare like "Night of the Living Dead" or "Hills Have Eyes." But for true horror fans, this one is worth checking out. Play it as the first entry on a double bill with the 1999 remake of "House on the Haunted Hill." The set-up and character dynamics are so similar that you really have to wonder what film they were actually remaking?



Text with highlighted words

'This movie took me by surprise. The opening credit sequence features nicely done animation. After that, we're plunged into a semi-cheesy production, betraying its low budget. The characters, typical American teens, are introduced slowly, with more personal detail than is usually found in movies like this. By the time the shlitz hits the fan, we know each one of the characters, and either like or hate them according to their distinct personalities. It's a slow uphill set-up, kind of like the ride up a slope of a really tall roller coaster. Thankfully, once the action kicks in, it's full blown old school HORROR! Steve Johnson's make-up effects are awesome. Equal in quality to much bigger budgeted films. And the scares are jolting. Kevin Tenney delivers his best movie ever, with heart-stopping surprises and creepy suspenseful set-ups. The tongue-in-cheek, sometimes cheesy, humor marks this film as pure 80s horror, as opposed to the sullen tone of earlier genre fare like "Night of the Living Dead" or "Hills Have Eyes." But for true horror fans, this one is worth checking out. Play it as the first entry on a double bill with the 1999 remake of "House on the Haunted Hill." The set-up and character dynamics are so similar that you really have to wonder what film they were actually remaking?

LIME with first sentence



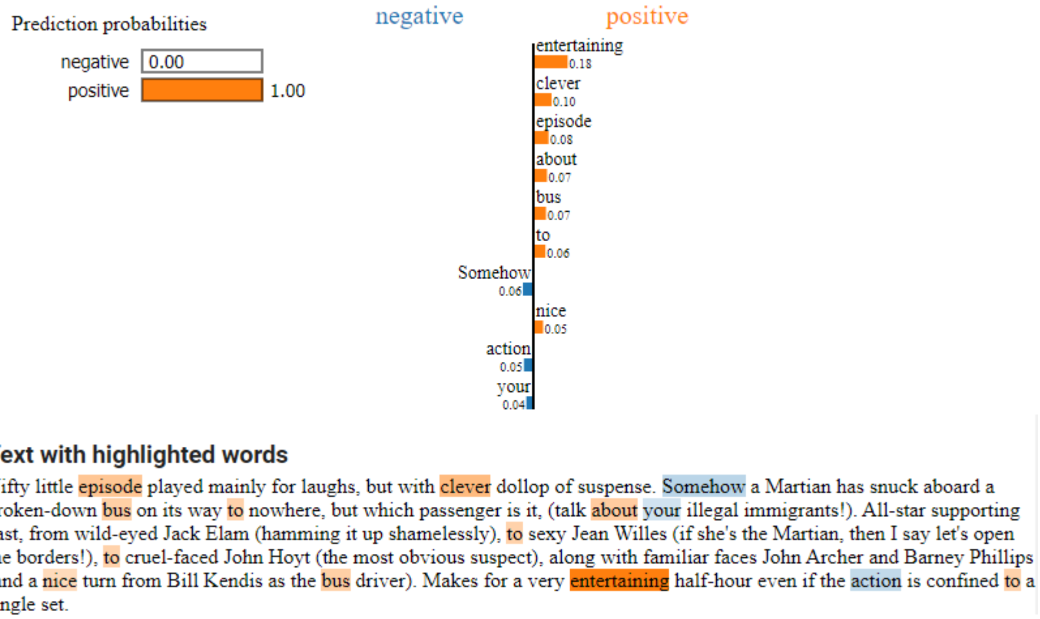
'This movie took me by surprise. The opening credit sequence features nicely done animation. After that, we're plunged into a semi-cheesy production, betraying its low budget. The characters, typical American teens, are introduced slowly, with more personal detail than is usually found in movies like this. By the time the shlitz hits the fan, we know each one of the characters, and either like or hate them according to their distinct personalities. It's a slow uphill set-up, kind of like the ride up a slope of a really tall roller coaster. Thankfully, once the action kicks in, it's full blown old school HORROR! Steve Johnson's make-up effects are awesome. Equal in quality to much bigger budgeted films. And the scares are jolting. Kevin Tenney delivers his best movie ever, with heart-stopping surprises and creepy suspenseful set-ups. The tongue-in-cheek, sometimes cheesy, humor marks this film as pure 80s horror, as opposed to the sullen tone of earlier genre fare like "Night of the Living Dead" or "Hills Have Eyes." But for true horror fans, this one is worth checking out. Play it as the first entry on a double bill with the 1999 remake of "House on the Haunted Hill." The set-up and character dynamics are so similar that you really have to wonder what film they were actually remaking?

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

SHAP with first sentence

2. Nifty little episode played mainly for laughs, but with clever dollop of suspense. Somehow a Martian has snuck aboard a broken-down bus on its way to

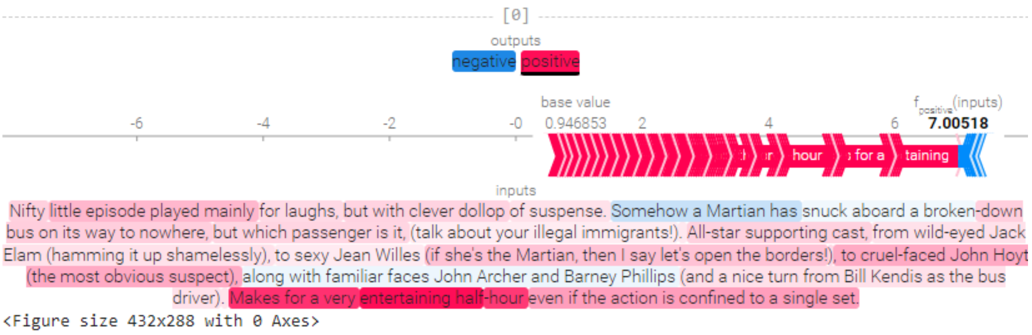
nowhere, but which passenger is it, (talk about your illegal immigrants!). All-star supporting cast, from wild-eyed Jack Elam (hamming it up shamelessly), to sexy Jean Willes (if she's the Martian, then I say let's open the borders!), to cruel-faced John Hoyt (the most obvious suspect), along with familiar faces John Archer and Barney Phillips (and a nice turn from Bill Kendis as the bus driver). Makes for a very entertaining half-hour even if the action is confined to a single set.



Text with highlighted words

Nifty little episode played mainly for laughs, but with clever dollop of suspense. Somehow a Martian has snuck aboard a broken-down bus on its way to nowhere, but which passenger is it, (talk about your illegal immigrants!). All-star supporting cast, from wild-eyed Jack Elam (hamming it up shamelessly), to sexy Jean Willes (if she's the Martian, then I say let's open the borders!), to cruel-faced John Hoyt (the most obvious suspect), along with familiar faces John Archer and Barney Phillips (and a nice turn from Bill Kendis as the bus driver). Makes for a very entertaining half-hour even if the action is confined to a single set.

LIME with second sentence



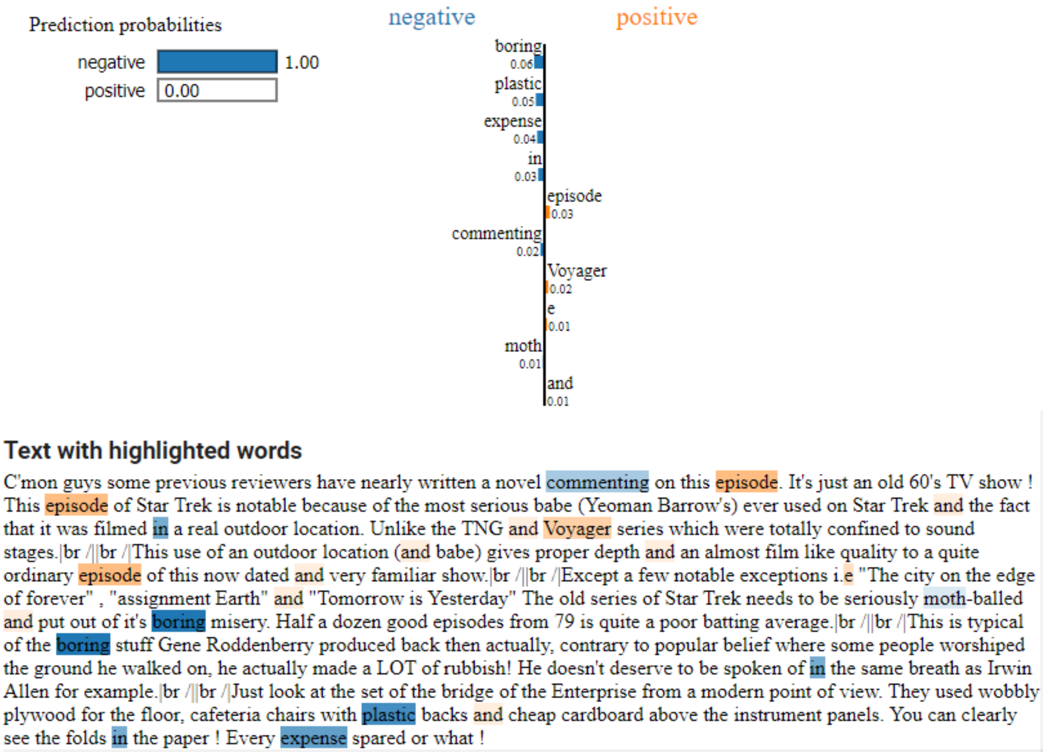
Nifty little episode played mainly for laughs, but with clever dollop of suspense. Somehow a Martian has snuck aboard a broken-down bus on its way to nowhere, but which passenger is it, (talk about your illegal immigrants!). All-star supporting cast, from wild-eyed Jack Elam (hamming it up shamelessly), to sexy Jean Willes (if she's the Martian, then I say let's open the borders!), to cruel-faced John Hoyt (the most obvious suspect), along with familiar faces John Archer and Barney Phillips (and a nice turn from Bill Kendis as the bus driver). Makes for a very entertaining half-hour even if the action is confined to a single set.

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

SHAP with second sentence

3. C'mon guys some previous reviewers have nearly written a novel commenting on this episode. It's just an old 60's TV show ! This episode of Star Trek is notable because of the most serious babe (Yeoman Barrow's) ever used on Star Trek and the fact that it was filmed in a real outdoor location. Unlike the TNG and Voyager series which were totally confined to sound stages.<br /><br />This use

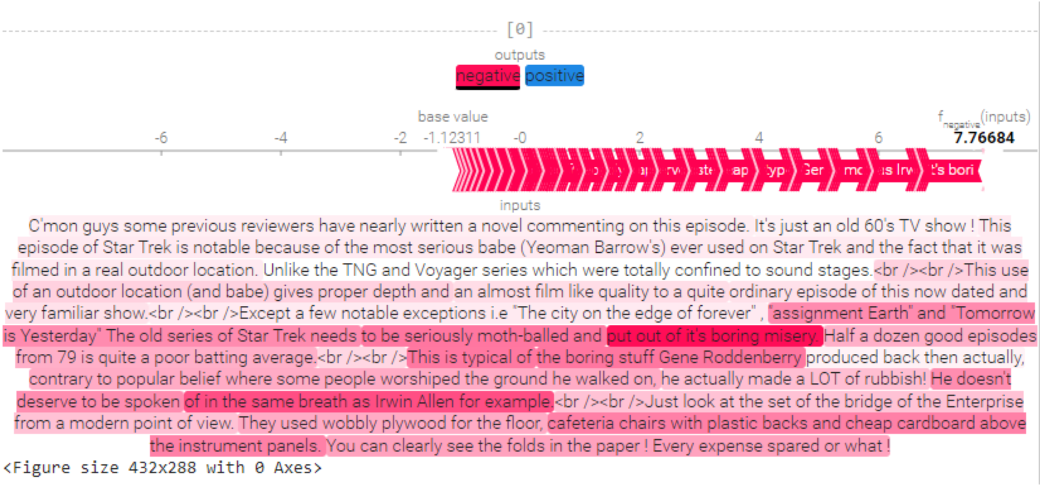
of an outdoor location (and babe) gives proper depth and an almost film like quality to a quite ordinary episode of this now dated and very familiar show.<br /><br />Except a few notable exceptions i.e "The city on the edge of forever" , "assignment Earth" and "Tomorrow is Yesterday" The old series of Star Trek needs to be seriously moth-balled and put out of it's boring misery. Half a dozen good episodes from 79 is quite a poor batting average.<br /><br />This is typical of the boring stuff Gene Roddenberry produced back then actually, contrary to popular belief where some people worshiped the ground he walked on, he actually made a LOT of rubbish! He doesn't deserve to be spoken of in the same breath as Irwin Allen for example.<br /><br />Just look at the set of the bridge of the Enterprise from a modern point of view. They used wobbly plywood for the floor, cafeteria chairs with plastic backs and cheap cardboard above the instrument panels. You can clearly see the folds in the paper ! Every expense spared or what !



Text with highlighted words

C'mon guys some previous reviewers have nearly written a novel commenting on this episode. It's just an old 60's TV show ! This episode of Star Trek is notable because of the most serious babe (Yeoman Barrow's) ever used on Star Trek and the fact that it was filmed in a real outdoor location. Unlike the TNG and Voyager series which were totally confined to sound stages.|br /||br /|This use of an outdoor location (and babe) gives proper depth and an almost film like quality to a quite ordinary episode of this now dated and very familiar show.|br /||br /|Except a few notable exceptions i.e "The city on the edge of forever" , "assignment Earth" and "Tomorrow is Yesterday" The old series of Star Trek needs to be seriously moth-balled and put out of it's boring misery. Half a dozen good episodes from 79 is quite a poor batting average.|br /||br /|This is typical of the boring stuff Gene Roddenberry produced back then actually, contrary to popular belief where some people worshiped the ground he walked on, he actually made a LOT of rubbish! He doesn't deserve to be spoken of in the same breath as Irwin Allen for example.|br /||br /|Just look at the set of the bridge of the Enterprise from a modern point of view. They used wobbly plywood for the floor, cafeteria chairs with plastic backs and cheap cardboard above the instrument panels. You can clearly see the folds in the paper ! Every expense spared or what !

LIME with third sentence

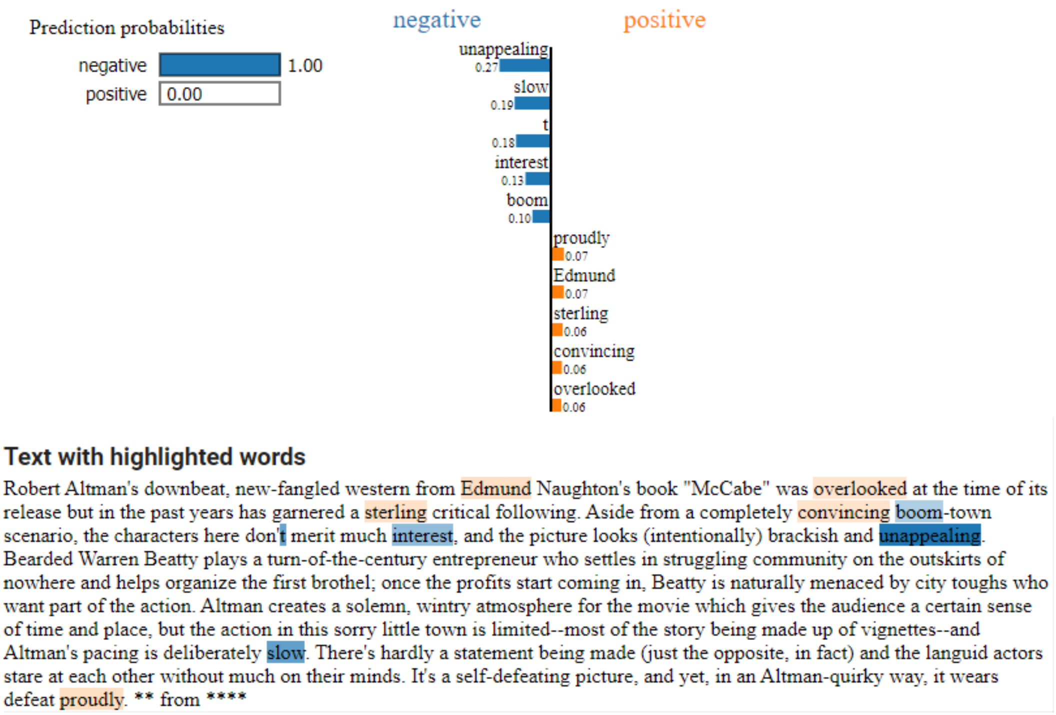


C'mon guys some previous reviewers have nearly written a novel commenting on this episode. It's just an old 60's TV show ! This episode of Star Trek is notable because of the most serious babe (Yeoman Barrow's) ever used on Star Trek and the fact that it was filmed in a real outdoor location. Unlike the TNG and Voyager series which were totally confined to sound stages.<br /><br />This use of an outdoor location (and babe) gives proper depth and an almost film like quality to a quite ordinary episode of this now dated and very familiar show.<br /><br />Except a few notable exceptions i.e "The city on the edge of forever" , "assignment Earth" and "Tomorrow is Yesterday" The old series of Star Trek needs to be seriously moth-balled and put out of it's boring misery. Half a dozen good episodes from 79 is quite a poor batting average.<br /><br />This is typical of the boring stuff Gene Roddenberry produced back then actually, contrary to popular belief where some people worshiped the ground he walked on, he actually made a LOT of rubbish! He doesn't deserve to be spoken of in the same breath as Irwin Allen for example.<br /><br />Just look at the set of the bridge of the Enterprise from a modern point of view. They used wobbly plywood for the floor, cafeteria chairs with plastic backs and cheap cardboard above the instrument panels. You can clearly see the folds in the paper ! Every expense spared or what !

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

SHAP with third sentence

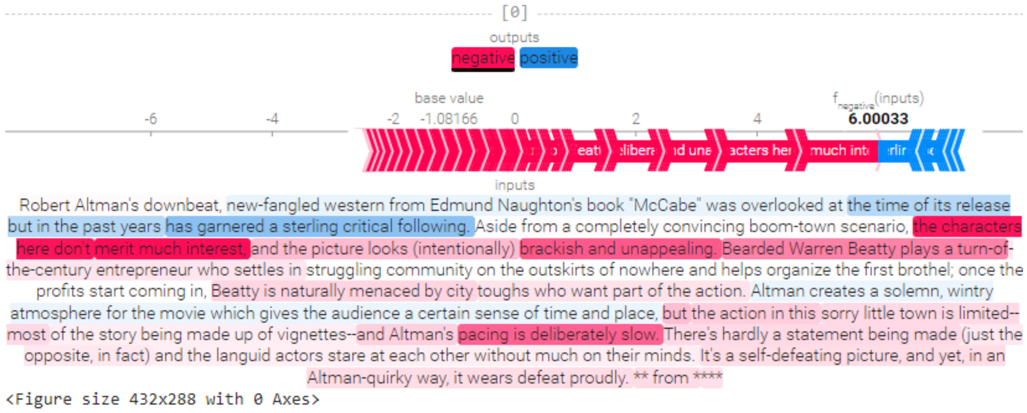
4. Robert Altman's downbeat, new-fangled western from Edmund Naughton's book "McCabe" was overlooked at the time of its release but in the past years has garnered a sterling critical following. Aside from a completely convincing boom-town scenario, the characters here don't merit much interest, and the picture looks (intentionally) brackish and unappealing. Bearded Warren Beatty plays a turn-of-the-century entrepreneur who settles in struggling community on the outskirts of nowhere and helps organize the first brothel; once the profits start coming in, Beatty is naturally menaced by city toughs who want part of the action. Altman creates a solemn, wintry atmosphere for the movie which gives the audience a certain sense of time and place, but the action in this sorry little town is limited--most of the story being made up of vignettes--and Altman's pacing is deliberately slow. There's hardly a statement being made (just the opposite, in fact) and the languid actors stare at each other without much on their minds. It's a self-defeating picture, and yet, in an Altman-quirky way, it wears defeat proudly. ** from ****



Text with highlighted words

Robert Altman's downbeat, new-fangled western from Edmund Naughton's book "McCabe" was overlooked at the time of its release but in the past years has garnered a sterling critical following. Aside from a completely convincing boom-town scenario, the characters here don't merit much interest, and the picture looks (intentionally) brackish and unappealing. Bearded Warren Beatty plays a turn-of-the-century entrepreneur who settles in struggling community on the outskirts of nowhere and helps organize the first brothel; once the profits start coming in, Beatty is naturally menaced by city toughs who want part of the action. Altman creates a solemn, wintry atmosphere for the movie which gives the audience a certain sense of time and place, but the action in this sorry little town is limited--most of the story being made up of vignettes--and Altman's pacing is deliberately slow. There's hardly a statement being made (just the opposite, in fact) and the languid actors stare at each other without much on their minds. It's a self-defeating picture, and yet, in an Altman-quirky way, it wears defeat proudly. ** from ****

LIME with fourth sentence



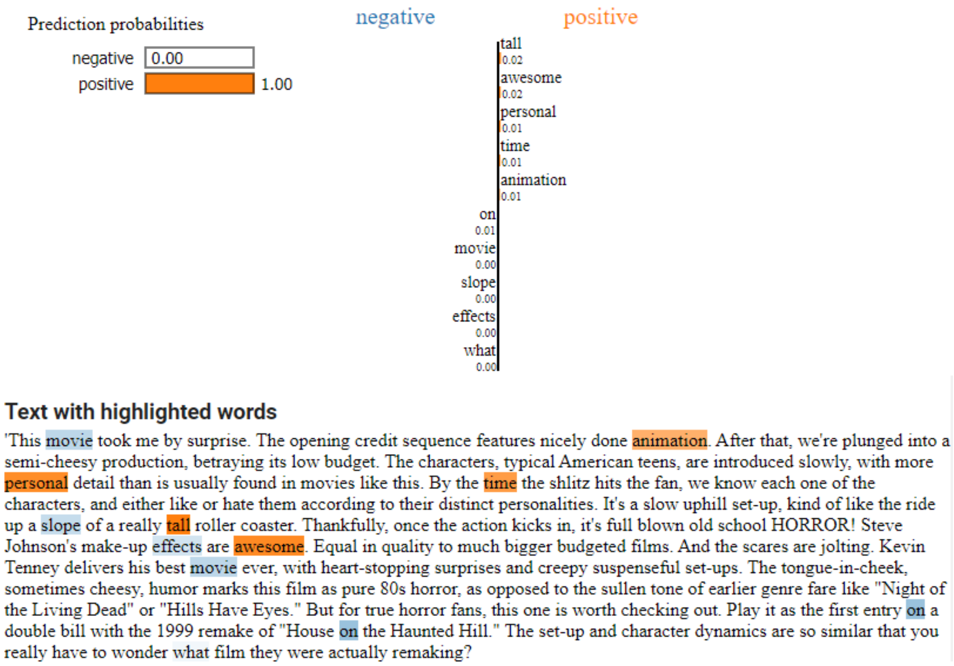
Robert Altman's downbeat, new-fangled western from Edmund Naughton's book "McCabe" was overlooked at the time of its release but in the past years has garnered a sterling critical following. Aside from a completely convincing boom-town scenario, the characters here don't merit much interest, and the picture looks (intentionally) brackish and unappealing. Bearded Warren Beatty plays a turn-of-the-century entrepreneur who settles in struggling community on the outskirts of nowhere and helps organize the first brothel; once the profits start coming in, Beatty is naturally menaced by city toughs who want part of the action. Altman creates a solemn, wintry atmosphere for the movie which gives the audience a certain sense of time and place, but the action in this sorry little town is limited--most of the story being made up of vignettes--and Altman's pacing is deliberately slow. There's hardly a statement being made (just the opposite, in fact) and the languid actors stare at each other without much on their minds. It's a self-defeating picture, and yet, in an Altman-quirky way, it wears defeat proudly. ** from ****

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

SHAP with fourth sentence

在 2.2 和 2.3 我們知道 LIME 和 SHAP 在簡短的句子裡都能有好的表現，而且當我們使用了一整個篇幅的電影評論時，兩者的表現也都是還不錯的，可是因為 LIME 一次只能標記一個單詞，而且每個單詞都會被視為一個特徵，所以如果遇到句子很長的情況，就會有過多特徵，導致很難訓練出一個簡單分類器來逼近原始模型，相反地，SHAP 則是能標記一整個句子，所以我認為在有些情況下，LIME 給出的解釋會不如 SHAP 給出的解釋，除此之外，在第三個例子中，LIME 將 episode 判斷為正面的詞語，這指出 LIME 的標記並不一定合理， 而 SHAP 給出的解釋明顯比較合理，因此，在情感分類的方面，我認為 SHAP 有著比 LIME 更好的解釋性，在 2.5 和 2.6 我將利用 SHAP 和 LIME 來比較助教所提供的兩個模型，分別為 distilbert-base-uncased 和 prajjwal1/bert-small，句子的部分則是使用例子一(正面)和例子三(負面)

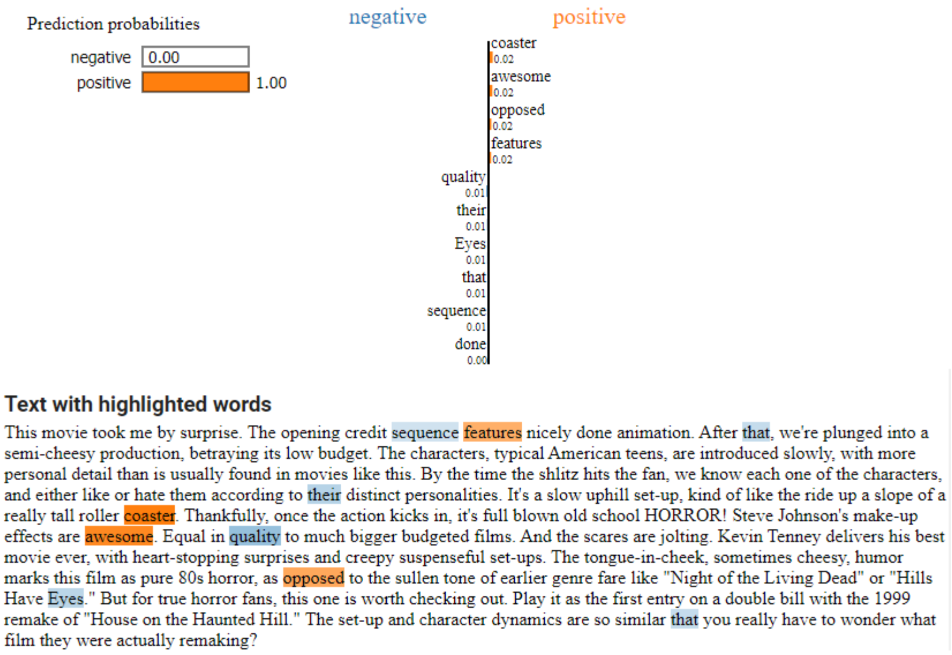
2.5 Comparison of two sentiment classification models with LIME



Text with highlighted words

'This movie took me by surprise. The opening credit sequence features nicely done animation. After that, we're plunged into a semi-cheesy production, betraying its low budget. The characters, typical American teens, are introduced slowly, with more personal detail than is usually found in movies like this. By the time the shlitz hits the fan, we know each one of the characters, and either like or hate them according to their distinct personalities. It's a slow uphill set-up, kind of like the ride up a slope of a really tall roller coaster. Thankfully, once the action kicks in, it's full blown old school HORROR! Steve Johnson's make-up effects are awesome. Equal in quality to much bigger budgeted films. And the scares are jolting. Kevin Tenney delivers his best movie ever, with heart-stopping surprises and creepy suspenseful set-ups. The tongue-in-cheek, sometimes cheesy, humor marks this film as pure 80s horror, as opposed to the sullen tone of earlier genre fare like "Night of the Living Dead" or "Hills Have Eyes." But for true horror fans, this one is worth checking out. Play it as the first entry on a double bill with the 1999 remake of "House on the Haunted Hill." The set-up and character dynamics are so similar that you really have to wonder what film they were actually remaking?

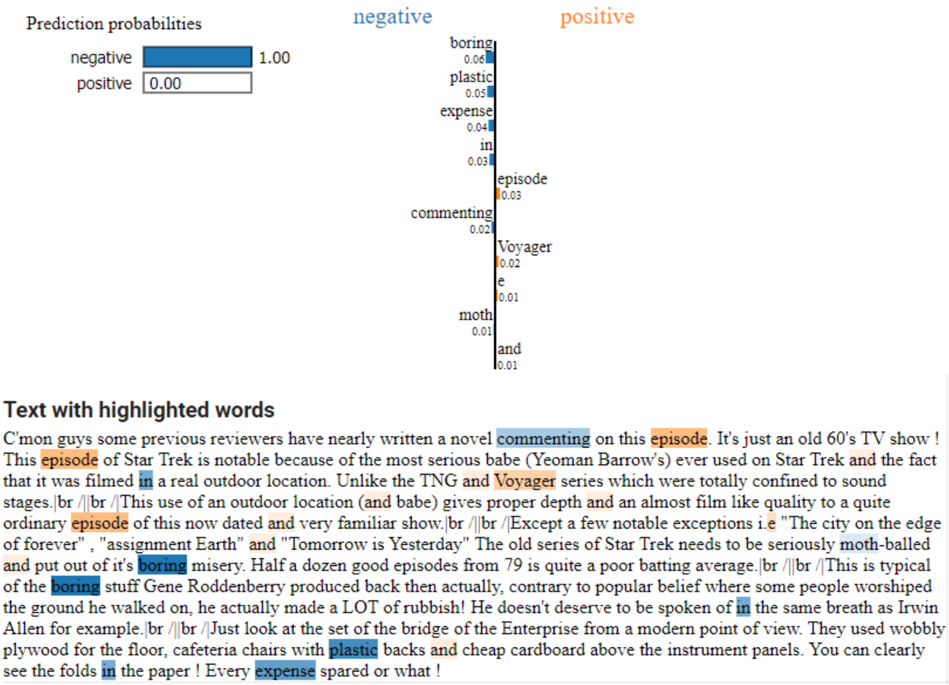
distilbert-base-uncased with first sentence



Text with highlighted words

This movie took me by surprise. The opening credit sequence features nicely done animation. After that, we're plunged into a semi-cheesy production, betraying its low budget. The characters, typical American teens, are introduced slowly, with more personal detail than is usually found in movies like this. By the time the shlitz hits the fan, we know each one of the characters, and either like or hate them according to their distinct personalities. It's a slow uphill set-up, kind of like the ride up a slope of a really tall roller coaster. Thankfully, once the action kicks in, it's full blown old school HORROR! Steve Johnson's make-up effects are awesome. Equal in quality to much bigger budgeted films. And the scares are jolting. Kevin Tenney delivers his best movie ever, with heart-stopping surprises and creepy suspenseful set-ups. The tongue-in-cheek, sometimes cheesy, humor marks this film as pure 80s horror, as opposed to the sullen tone of earlier genre fare like "Night of the Living Dead" or "Hills Have Eyes." But for true horror fans, this one is worth checking out. Play it as the first entry on a double bill with the 1999 remake of "House on the Haunted Hill." The set-up and character dynamics are so similar that you really have to wonder what film they were actually remaking?

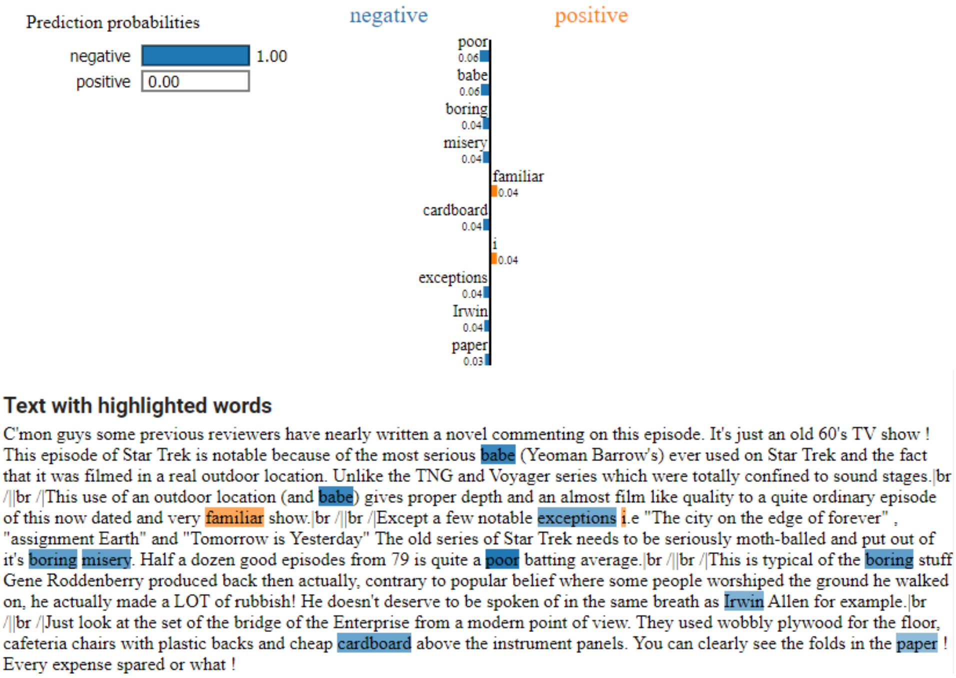
prajjwal1/bert-small with third sentence



Text with highlighted words

C'mon guys some previous reviewers have nearly written a novel commenting on this episode. It's just an old 60's TV show ! This episode of Star Trek is notable because of the most serious babe (Yeoman Barrow's) ever used on Star Trek and the fact that it was filmed in a real outdoor location. Unlike the TNG and Voyager series which were totally confined to sound stages.|br /||br /|This use of an outdoor location (and babe) gives proper depth and an almost film like quality to a quite ordinary episode of this now dated and very familiar show.|br /||br /|Except a few notable exceptions i.e "The city on the edge of forever" , "assignment Earth" and "Tomorrow is Yesterday" The old series of Star Trek needs to be seriously moth-balled and put out of it's boring misery. Half a dozen good episodes from 79 is quite a poor batting average.|br /||br /|This is typical of the boring stuff Gene Roddenberry produced back then actually, contrary to popular belief where some people worshiped the ground he walked on, he actually made a LOT of rubbish! He doesn't deserve to be spoken of in the same breath as Irwin Allen for example.|br /||br /|Just look at the set of the bridge of the Enterprise from a modern point of view. They used wobbly plywood for the floor, cafeteria chairs with plastic backs and cheap cardboard above the instrument panels. You can clearly see the folds in the paper ! Every expense spared or what !

distilbert-base-uncased with third sentence



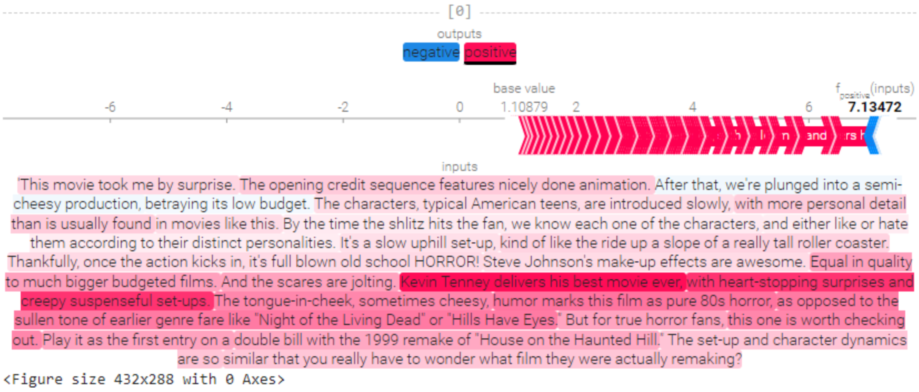
Text with highlighted words

C'mon guys some previous reviewers have nearly written a novel commenting on this episode. It's just an old 60's TV show ! This episode of Star Trek is notable because of the most serious babe (Yeoman Barrow's) ever used on Star Trek and the fact that it was filmed in a real outdoor location. Unlike the TNG and Voyager series which were totally confined to sound stages.|br /||br /|This use of an outdoor location (and babe) gives proper depth and an almost film like quality to a quite ordinary episode of this now dated and very familiar show.|br /||br /|Except a few notable exceptions i.e "The city on the edge of forever" , "assignment Earth" and "Tomorrow is Yesterday" The old series of Star Trek needs to be seriously moth-balled and put out of it's boring misery. Half a dozen good episodes from 79 is quite a poor batting average.|br /||br /|This is typical of the boring stuff Gene Roddenberry produced back then actually, contrary to popular belief where some people worshiped the ground he walked on, he actually made a LOT of rubbish! He doesn't deserve to be spoken of in the same breath as Irwin Allen for example.|br /||br /|Just look at the set of the bridge of the Enterprise from a modern point of view. They used wobbly plywood for the floor, cafeteria chairs with plastic backs and cheap cardboard above the instrument panels. You can clearly see the folds in the paper ! Every expense spared or what !

prajjwal1/bert-small with third sentence

在上述兩個例子可以發現，有許多不帶有情緒的詞語都會被 LIME 標記出來，像是例子一中distilbert-base-uncased 所標記出的 on 還有 prajjwal1/bert-small 所標記出的 their 和 that，而且兩個模型所標記的詞語也十分不一致，甚至，例子三的 poor 明顯為負面詞語，在 distilbert-base-uncased 卻沒被抓出來，所以我認為無論是哪個模型，套用在 LIME 上似乎都沒有很好的解釋性。

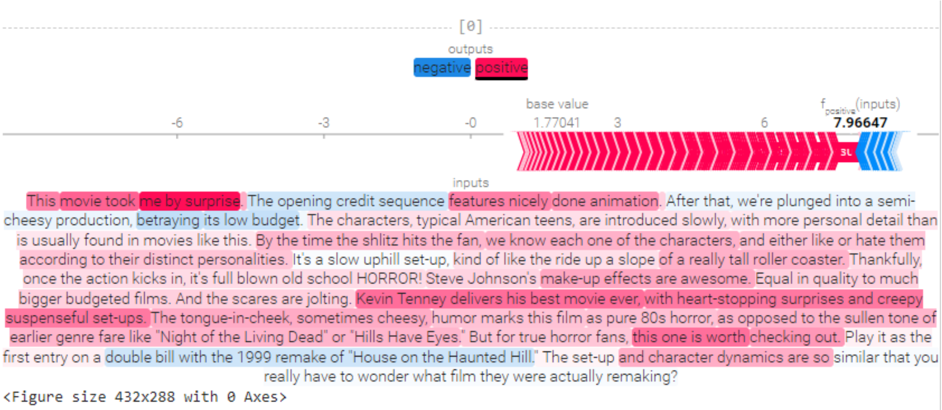
2.6 Comparison of two sentiment classification models with SHAP



'This movie took me by surprise. The opening credit sequence features nicely done animation. After that, we're plunged into a semi-cheesy production, betraying its low budget. The characters, typical American teens, are introduced slowly, with more personal detail than is usually found in movies like this. By the time the shlitz hits the fan, we know each one of the characters, and either like or hate them according to their distinct personalities. It's a slow uphill set-up, kind of like the ride up a slope of a really tall roller coaster. Thankfully, once the action kicks in, it's full blown old school HORROR! Steve Johnson's make-up effects are awesome. Equal in quality to much bigger budgeted films. And the scares are jolting. Kevin Tenney delivers his best movie ever, with heart-stopping surprises and creepy suspenseful set-ups. The tongue-in-cheek, sometimes cheesy, humor marks this film as pure 80s horror, as opposed to the sullen tone of earlier genre fare like "Night of the Living Dead" or "Hills Have Eyes." But for true horror fans, this one is worth checking out. Play it as the first entry on a double bill with the 1999 remake of "House on the Haunted Hill." The set-up and character dynamics are so similar that you really have to wonder what film they were actually remaking?

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

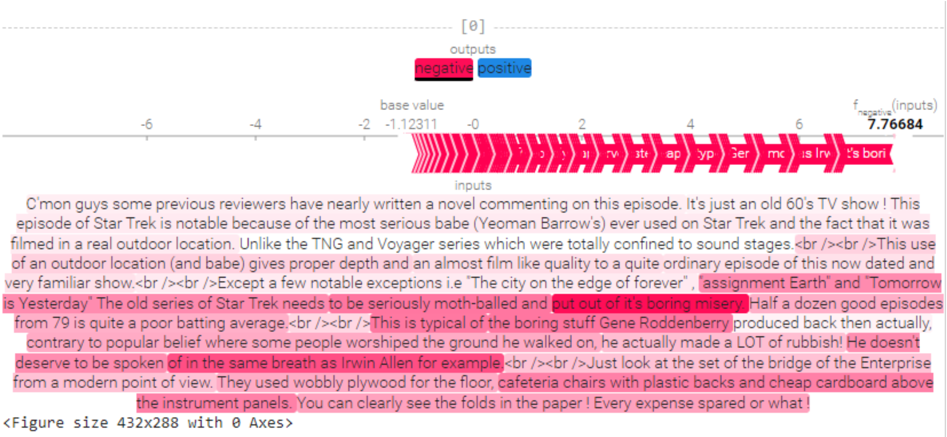
distilbert-base-uncased with first sentence



This movie took me by surprise. The opening credit sequence features nicely done animation. After that, we're plunged into a semi-cheesy production, betraying its low budget. The characters, typical American teens, are introduced slowly, with more personal detail than is usually found in movies like this. By the time the shlitz hits the fan, we know each one of the characters, and either like or hate them according to their distinct personalities. It's a slow uphill set-up, kind of like the ride up a slope of a really tall roller coaster. Thankfully, once the action kicks in, it's full blown old school HORROR! Steve Johnson's make-up effects are awesome. Equal in quality to much bigger budgeted films. And the scares are jolting. Kevin Tenney delivers his best movie ever, with heart-stopping surprises and creepy suspenseful set-ups. The tongue-in-cheek, sometimes cheesy, humor marks this film as pure 80s horror, as opposed to the sullen tone of earlier genre fare like "Night of the Living Dead" or "Hills Have Eyes." But for true horror fans, this one is worth checking out. Play it as the first entry on a double bill with the 1999 remake of "House on the Haunted Hill." The set-up and character dynamics are so similar that you really have to wonder what film they were actually remaking?

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

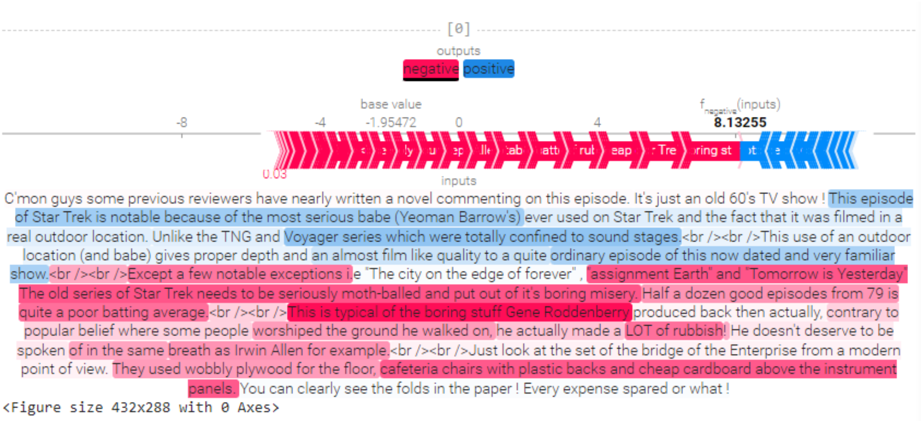
distilbert-base-uncased with first sentence



C'mon guys some previous reviewers have nearly written a novel commenting on this episode. It's just an old 60's TV show ! This episode of Star Trek is notable because of the most serious babe (Yeoman Barrow's) ever used on Star Trek and the fact that it was filmed in a real outdoor location. Unlike the TNG and Voyager series which were totally confined to sound stages.<br /><br />This use of an outdoor location (and babe) gives proper depth and an almost film like quality to a quite ordinary episode of this now dated and very familiar show.<br /><br />Except a few notable exceptions i.e "The city on the edge of forever" , "assignment Earth" and "Tomorrow is Yesterday" The old series of Star Trek needs to be seriously moth-balled and put out of it's boring misery. Half a dozen good episodes from 79 is quite a poor batting average.<br /><br />This is typical of the boring stuff Gene Roddenberry produced back then actually, contrary to popular belief where some people worshiped the ground he walked on, he actually made a LOT of rubbish! He doesn't deserve to be spoken of in the same breath as Irwin Allen for example.<br /><br />Just look at the set of the bridge of the Enterprise from a modern point of view. They used wobbly plywood for the floor, cafeteria chairs with plastic backs and cheap cardboard above the instrument panels. You can clearly see the folds in the paper ! Every expense spared or what !

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

distilbert-base-uncased with third sentence



C'mon guys some previous reviewers have nearly written a novel commenting on this episode. It's just an old 60's TV show ! This episode of Star Trek is notable because of the most serious babe (Yeoman Barrow's) ever used on Star Trek and the fact that it was filmed in a real outdoor location. Unlike the TNG and Voyager series which were totally confined to sound stages.<br /><br />This use of an outdoor location (and babe) gives proper depth and an almost film like quality to a quite ordinary episode of this now dated and very familiar show.<br /><br />Except a few notable exceptions i.e "The city on the edge of forever" , "assignment Earth" and "Tomorrow is Yesterday" The old series of Star Trek needs to be seriously moth-balled and put out of it's boring misery. Half a dozen good episodes from 79 is quite a poor batting average.<br /><br />This is typical of the boring stuff Gene Roddenberry produced back then actually, contrary to popular belief where some people worshiped the ground he walked on, he actually made a LOT of rubbish! He doesn't deserve to be spoken of in the same breath as Irwin Allen for example.<br /><br />Just look at the set of the bridge of the Enterprise from a modern point of view. They used wobbly plywood for the floor, cafeteria chairs with plastic backs and cheap cardboard above the instrument panels. You can clearly see the folds in the paper ! Every expense spared or what !

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

prajjwal1/bert-small with third sentence

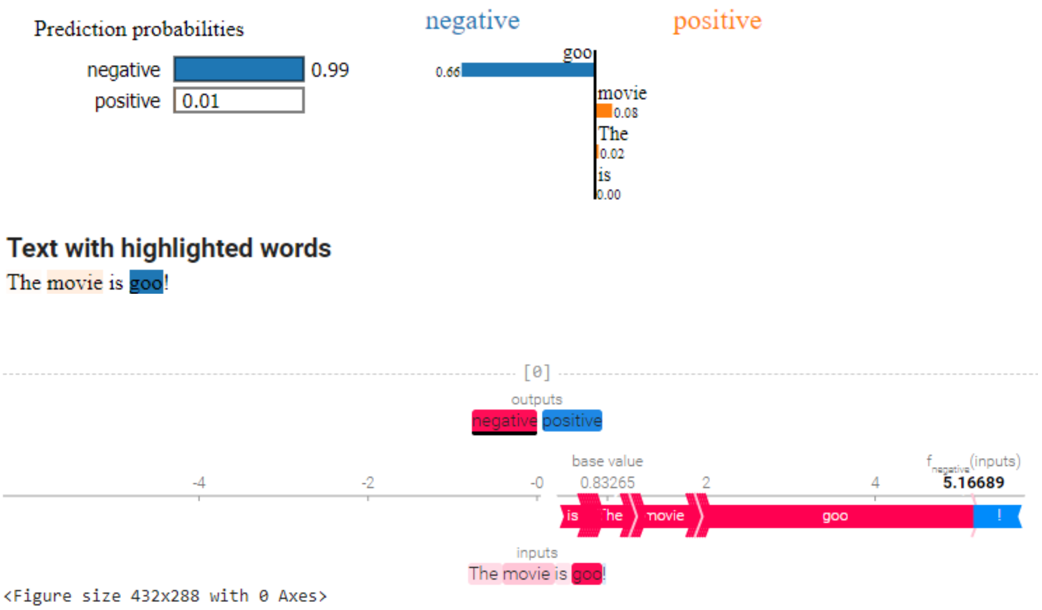
在前兩張圖可以發現，有兩個句子在兩個模型間的判斷並不相同，然而那兩個句子並沒有好壞之分，這並沒有關係，但在第三個例子裡，prajjwal1/bert-small 卻將一些我認為是正向的語句判斷為負面的，所以我認為 distilbert-base-uncased 的判斷似乎較為準確一些，但整體來看，兩者並沒有太大的差別。

Attack in NLP

在這個部份，我選擇攻擊助教所提供的第一個模型 (distilbert-base-uncased)。

3.1 Misspelling

在拼錯字的情況下，確實有可能導致判斷錯誤，舉例來說，假設我不小心將 good 拼成 god，原本應為正面的一個評論，卻被判定為負面的，因此，這是一個成功的攻擊。



Text with highlighted words

The movie is goo!

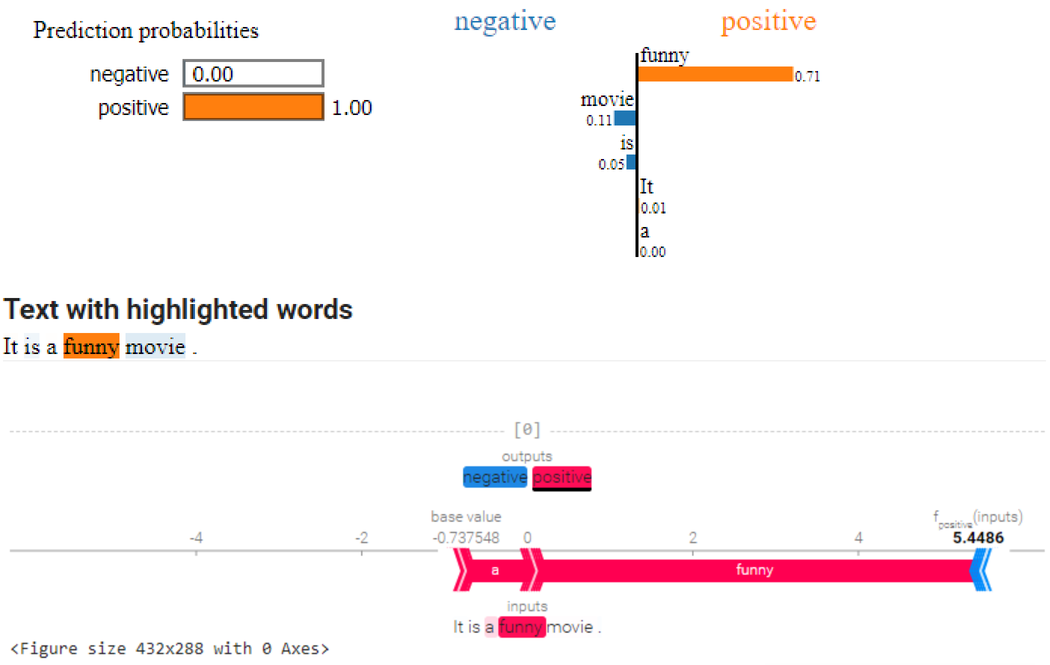
The movie is goo!

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

using the sentence “The movie is goo”

3.2 Word deletion

直接刪掉一個詞語應該是一個很簡單的攻擊，如果我刪掉的是句子裡的關鍵詞語，像是以下的 funny，導致句子的結構和語義被破壞，這樣模型便有極高的機率判斷錯誤，但缺點是馬上就會被發現了。

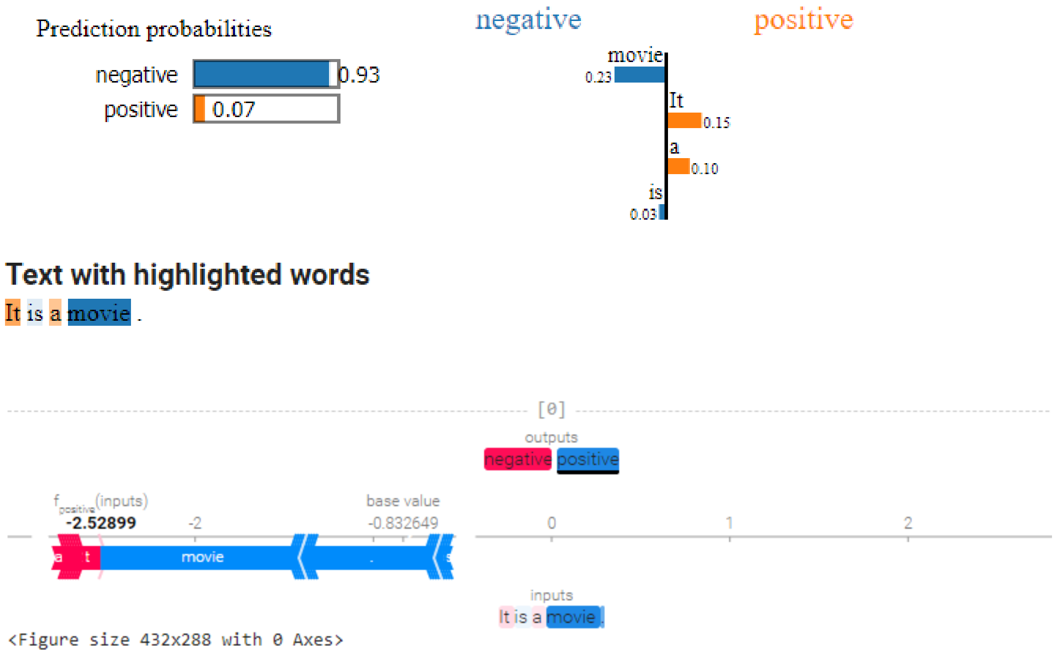


Text with highlighted words

It is a funny movie .

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

using the sentence “It is a funny movie.”



Text with highlighted words

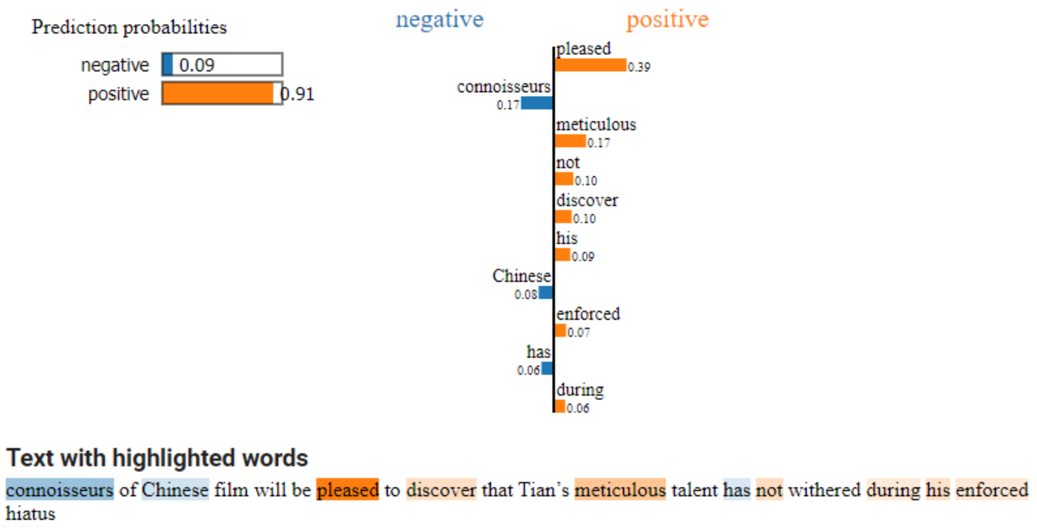
It is a movie .

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

using the sentence “It is a movie.”

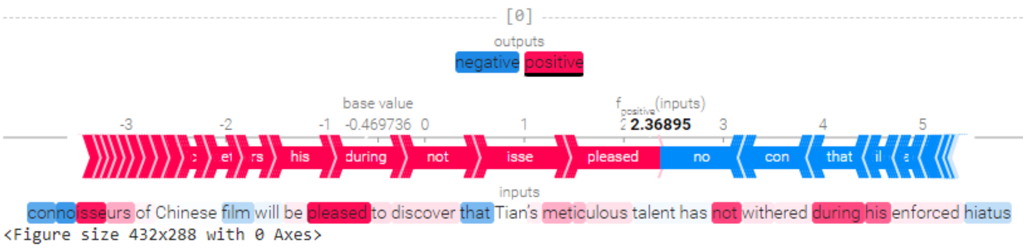
3.3 Word substitution by synonym

在這個例子裡，我將 film 這個字用 footage 替代，整句話的語意基本上仍是一樣的，但在第二張圖片中可以看到，footage 被歸類在負面的詞語，導致正面的分數只剩下 61%，如果再多替換幾個同義字，或許句子的判斷就會被歸類為負面的。



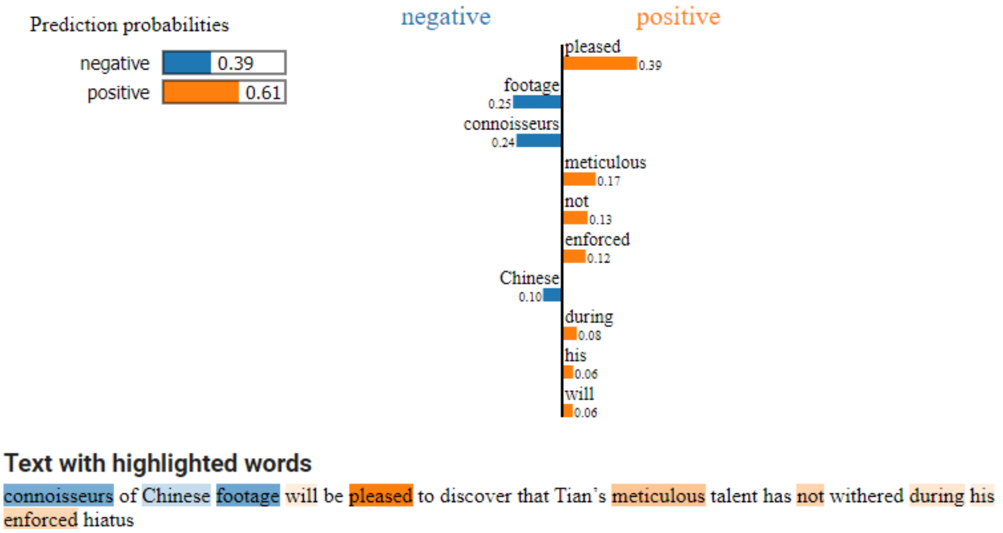
Text with highlighted words

connoisseurs of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus



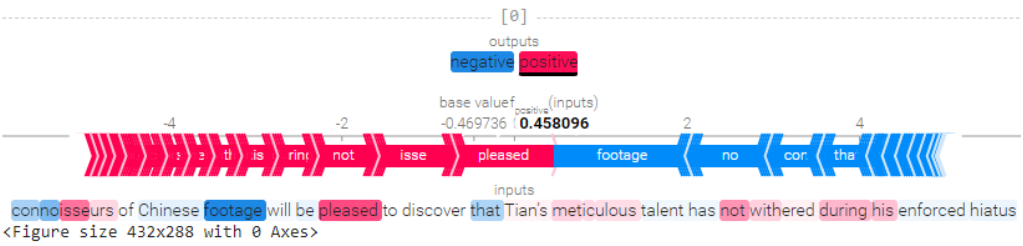
```
<Figure size 432x288 with 0 Axes>
```

using the sentence “connoisseurs of Chinese film will be pleased to discover that Tian’s meticulous talent has not withered during his enforced hiatus”



Text with highlighted words

connoisseurs of Chinese footage will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus

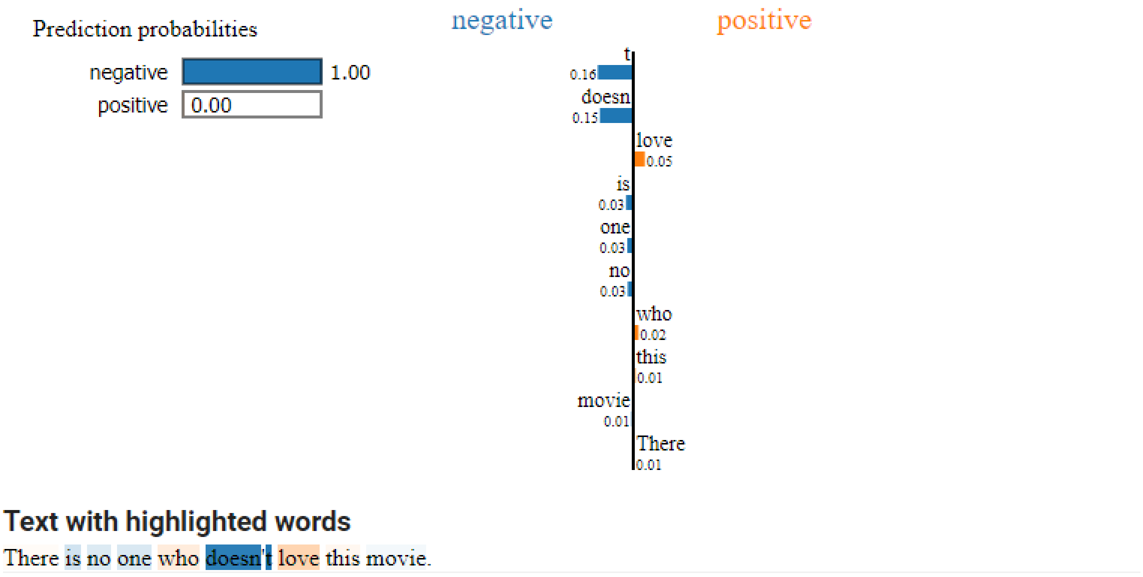


```
<Figure size 432x288 with 0 Axes>
```

using the sentence “connoisseurs of Chinese footage will be pleased to discover that Tian’s meticulous talent has not withered during his enforced hiatus”

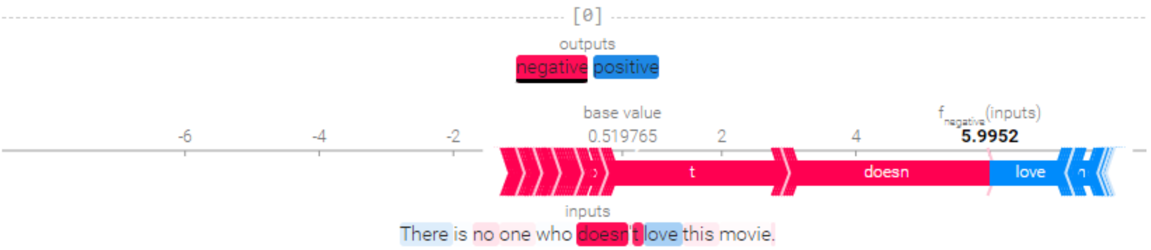
3.4 Put negative words in the sentence with positive sentiment

我試著在句子裡放入很多帶有負面情緒的詞語，但如果看一整個句子，會發現句子實際上是帶有正面意義的，舉例來說，下方第一個例子是在說大家都很喜歡看這部電影，實際上是個極高的評價，卻因為帶有 doesn’t, no 這些負面字眼，反而被判斷成負面評論，還有第二個例子，原句說明了該電影很感人，但也因為句中帶有 never 這種負面詞語，導致判斷結果不如預期。



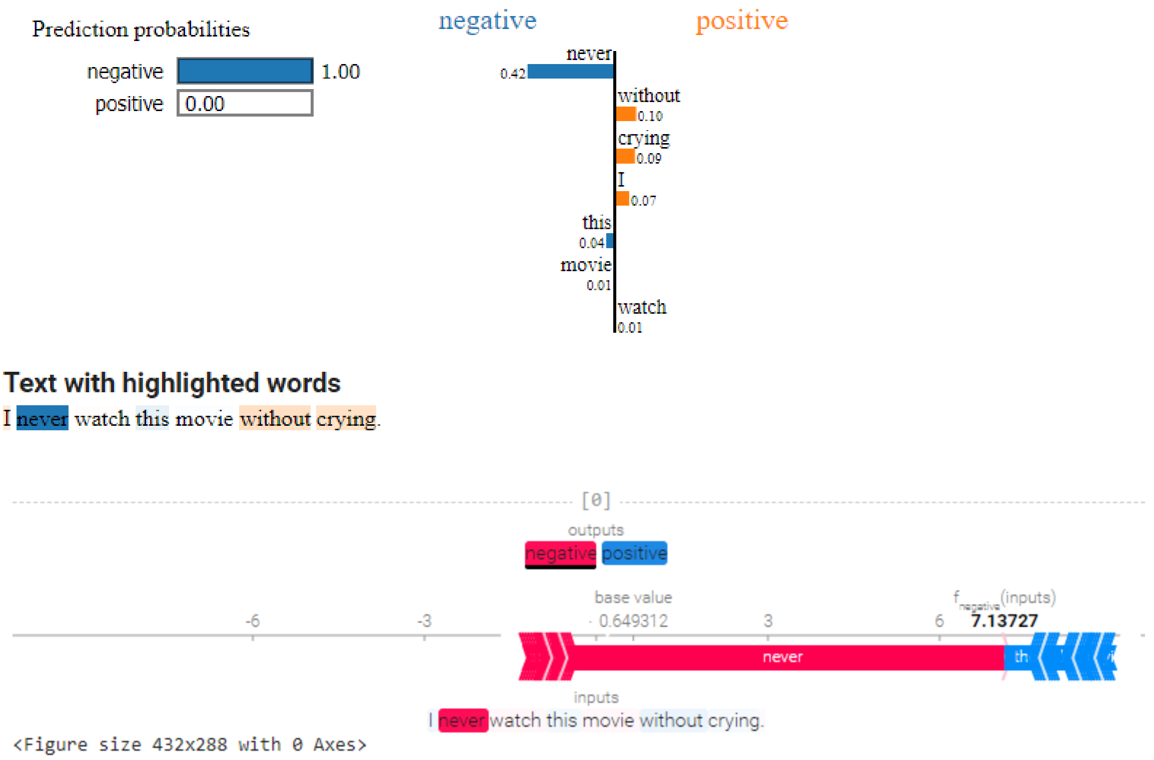
Text with highlighted words

There is no one who doesn't love this movie.



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

using the sentence “There is no one who doesn't love this movie.”



I never watch this movie without crying.

I never watch this movie without crying.

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

using the sentence “I never watch this movie without crying.”

3.5 Prevention

那些對語意影響較大的詞語便是易遭受攻擊的詞語，所以我們可以試著將每個詞遮住，看哪個詞被遮住後，對分數的影響最大，那該詞便是易遭受攻擊的詞語，接著我們可以使用同義詞詞典過濾反義詞，將那些詞語替換掉，並將替換完成的句子重新輸入，如果輸出和原先相反的判斷，那麼輸出該句子就可以被作為對抗樣本，以抵擋那些攻擊。

Problems

1. 一開始我有點不確定該選擇 exBERT 上的第幾層，後來透過一些觀察，才慢慢看出一個所以然，開始知道該如何選擇，並且將不同的模型進行比較。
2. 在使用 IMDB 中的評論的時候，句子中會有 we’re 或是 “McCabe“ 這些有單引號或是雙引號的詞語，這時就須使用三引號去包夾句子，不然會受影響。
3. NLP 攻擊的方式有很多種，但我一直找不到既能不改變語義又能改變判斷結果的方法，唯一找到最接近的應該就是 3.3 的結果了，雖然說語義確實沒被改變，但判斷結果沒達到我想要的結果，即便如此，但我還是認為應該有其他方法能達到我預想的條件。