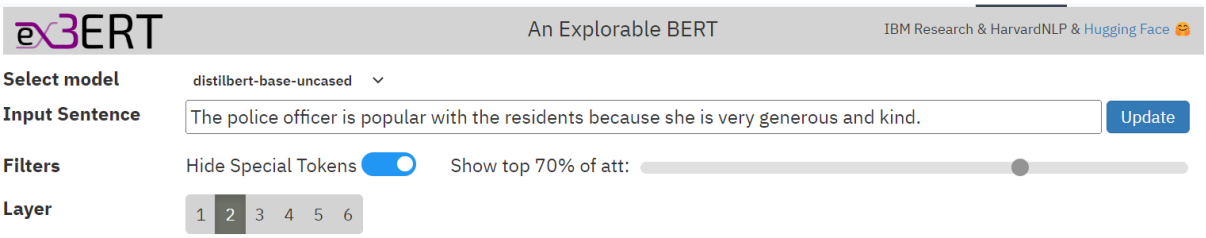
hw4_111511198

📌 Author : 111511198 Heng-An Cheng

Part 1 Attention Visualization - exBERT

Input Sentence :

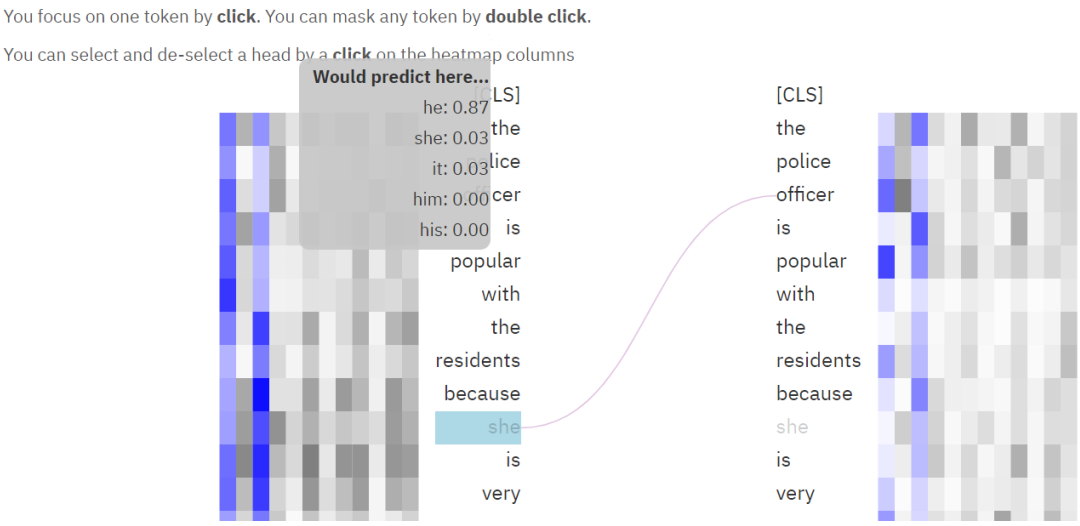
“The police officer is popular with the residents because she is very generous and kind.”



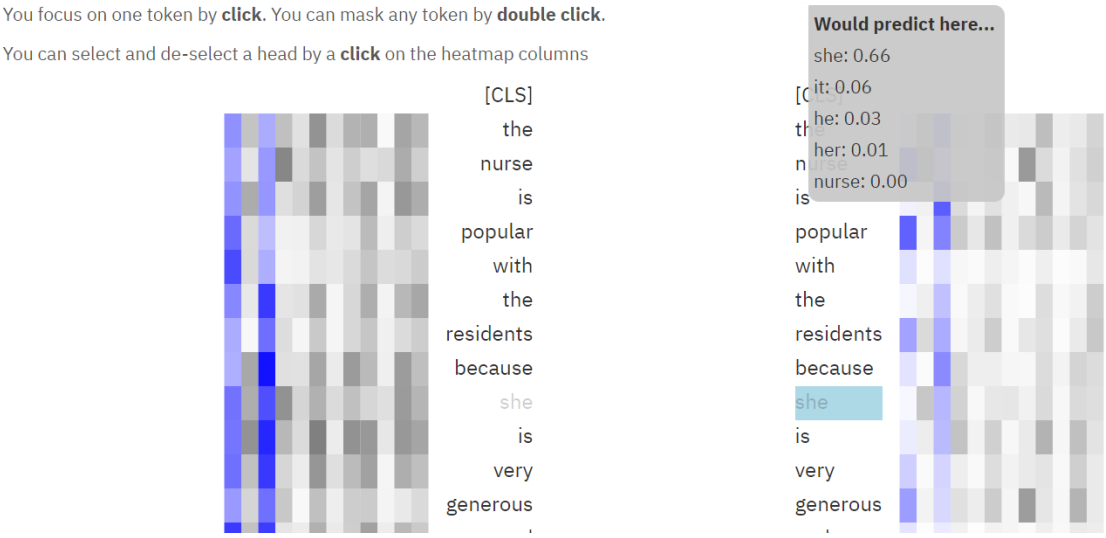
Masked word

Discover

If we mask the word she, the model will perdict the masked place he with 0.87 and only 0.03 for she. I think maybe it’s prediction has gender sterotype to the word “ police officer” at the beginning of the sentence.



So I change the “police officer” into “nurse”, we can find that the result is totally different, with 0.66 would predict she



Heads and Layer

Discover

Upon comparing head 4 across different layers, several patterns can be observed:

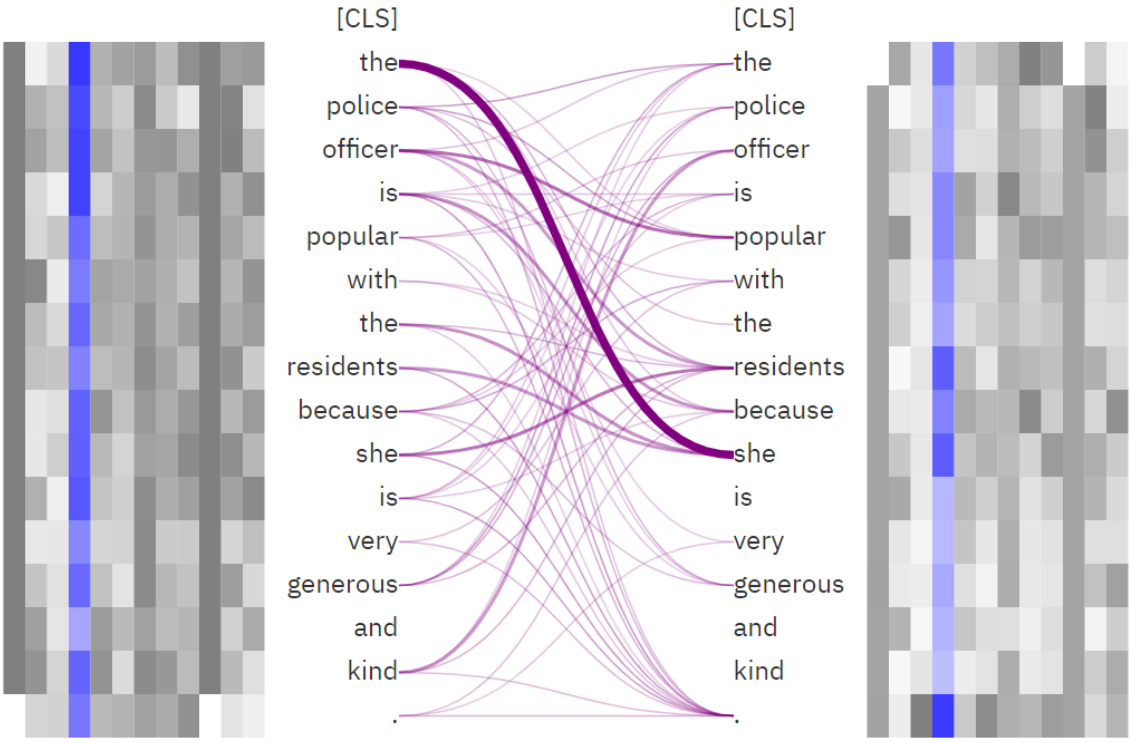
1. In Layer 1, head 4 connects the word before the target word in the sentence, suggesting a relationship or dependency between them.

2. In Layer 2, head 4 connects words that seem to have a relation or connection within the sentence.
3. In Layer 3, head 4 connects identical words within the sentence, indicating a self-connection or repetition.
4. In Layer 4, head 4 only connects with the words "and" and "popular," which might imply that these words are crucial or influential for the model's predictions. This suggests that this head might capture important features.
5. In Layer 5, head 4 connects words with similar meanings or related concepts within the sentence. For example, it might connect "police officer" and "she" if they are referring to the same entity.
6. In Layer 6, head 4 connects all words in the sentence to the punctuation mark ".", indicating a broader association between the words and the sentence-ending punctuation.

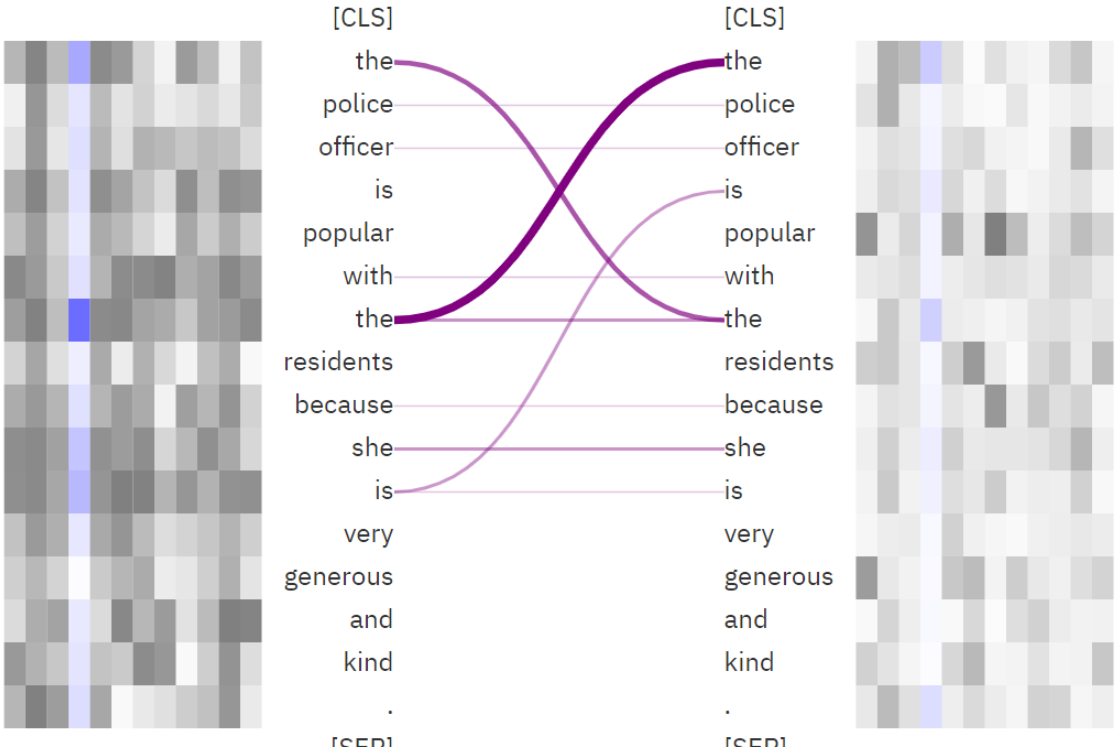
These observations demonstrate the different roles and relationships captured by head 4 in each layer, highlighting the complex and multi-dimensional nature of the model's attention patterns.

Layer 1

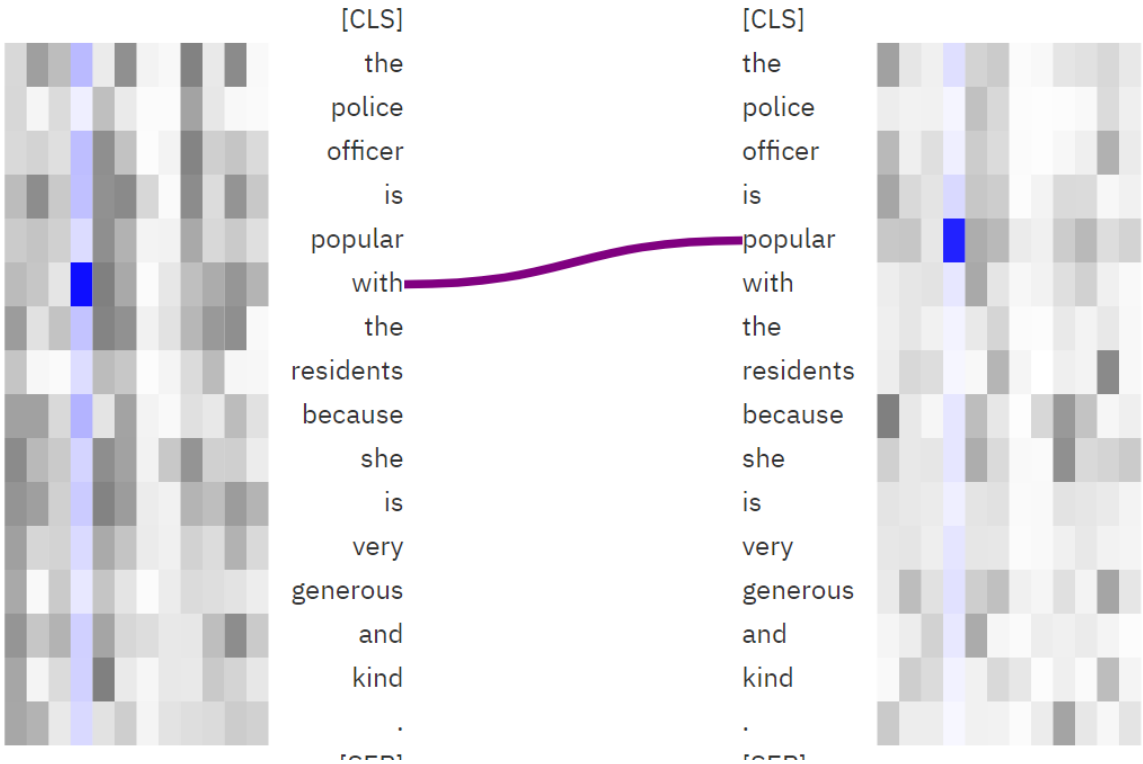
Layer 2



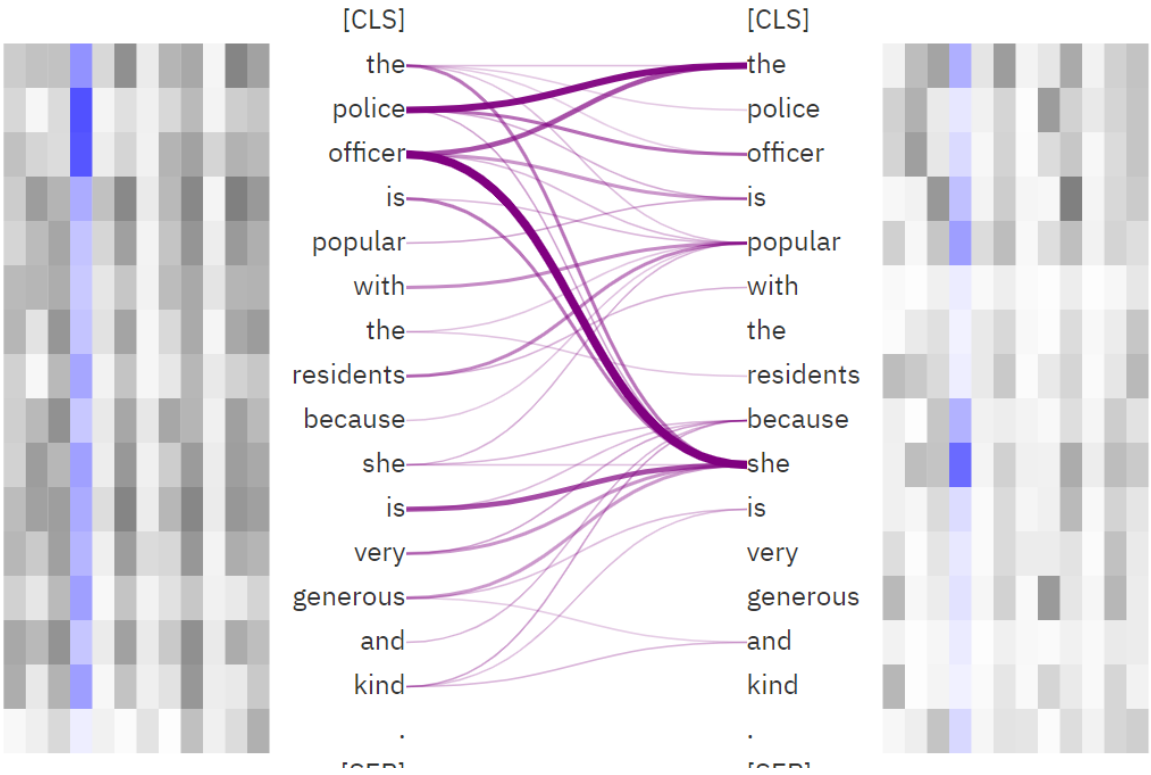
Layer 3



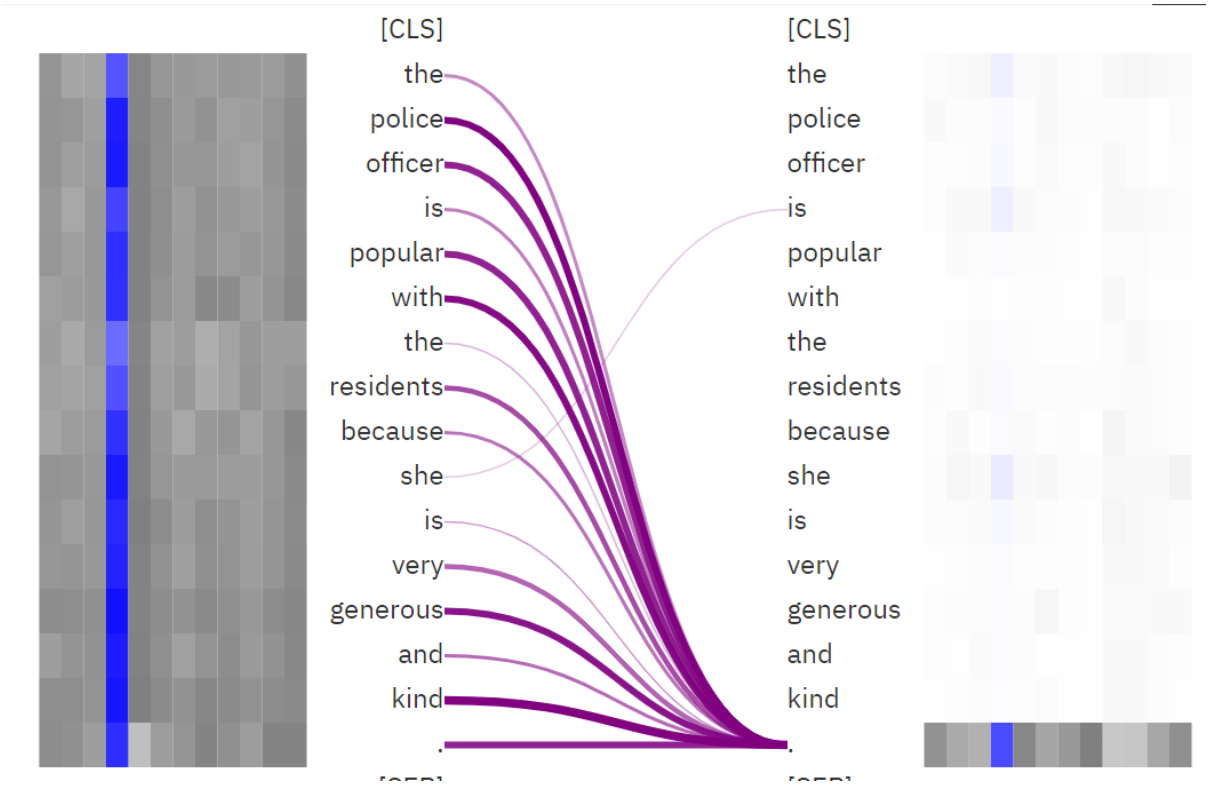
Layer 4



Layer 5



Layer 6



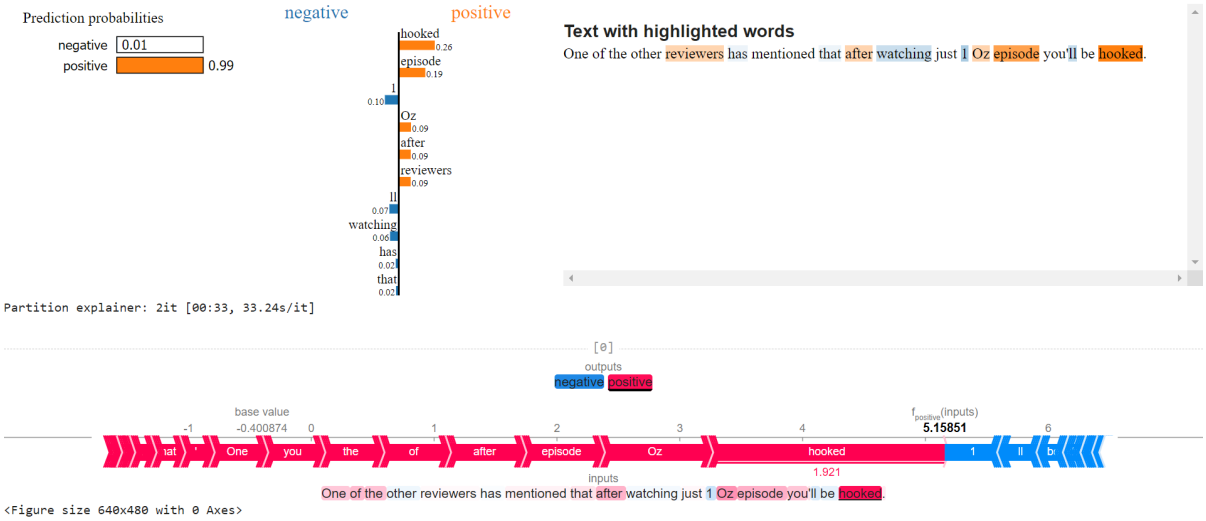
Part 2 Explanation Techniques - LIME & SHAP

Ex1:

"One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked."

Discover

As we can see from the chart below, the prediction and the reason of the result are almost the same from LIME and SHAP method. But we can find that, even the word “hooked” are viewed as the important word to determine the result, it only consists of 0.26 out of 1 in the former one, while 0.37(1.921/5.159) out of 1 in SHAP.

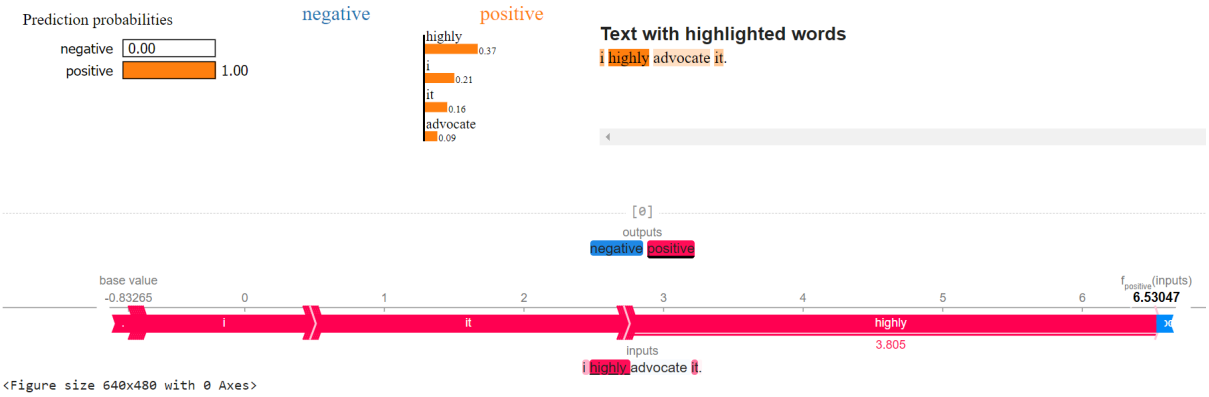


Ex2: "I high advocate it.”

Discover

As we can learn from the chart below, the prediction and the reason of the result are almost the same from LIME and SHAP method. And the word “highly” indeed makes the sentence more positive with 0.37/1 and 3.81/6.53 respectively in TA_model_1.

TA_model_1

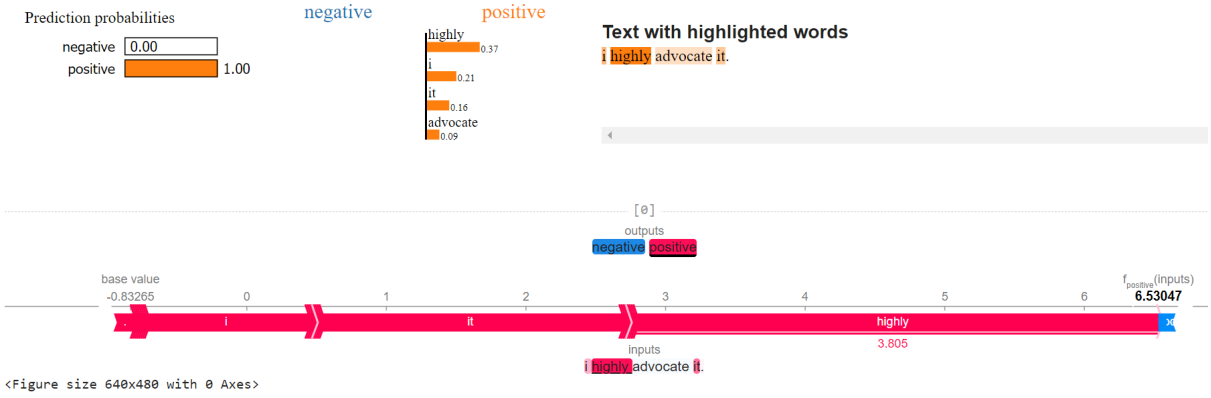


Part 3 Explanation Techniques on different model

Positive prediction - “I highly advocate it.”

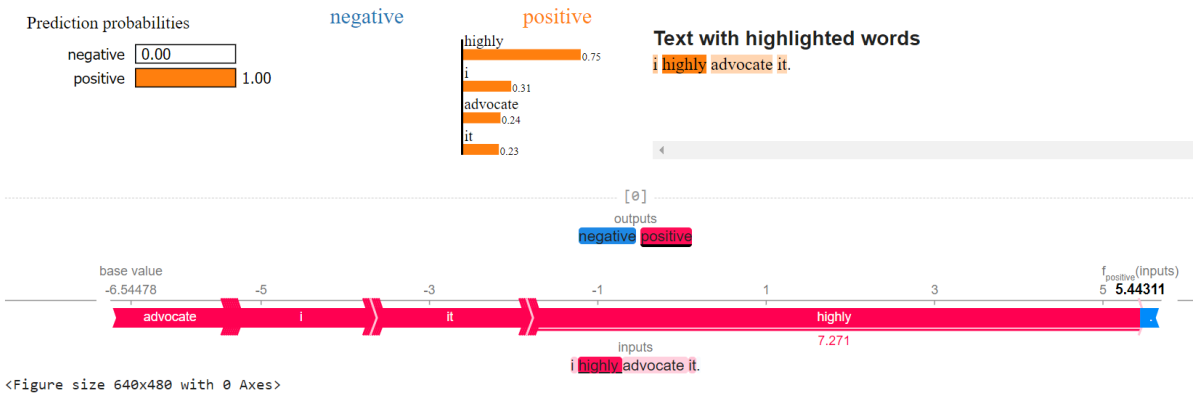
We can observe that although both models predict a positive outcome, the reasons behind the predictions differ. For instance, the word "highly" is assigned different levels of importance in the two models. In TA_model_2, it contributes a higher score than in the other model.

TA_model_1




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

TA_model_2

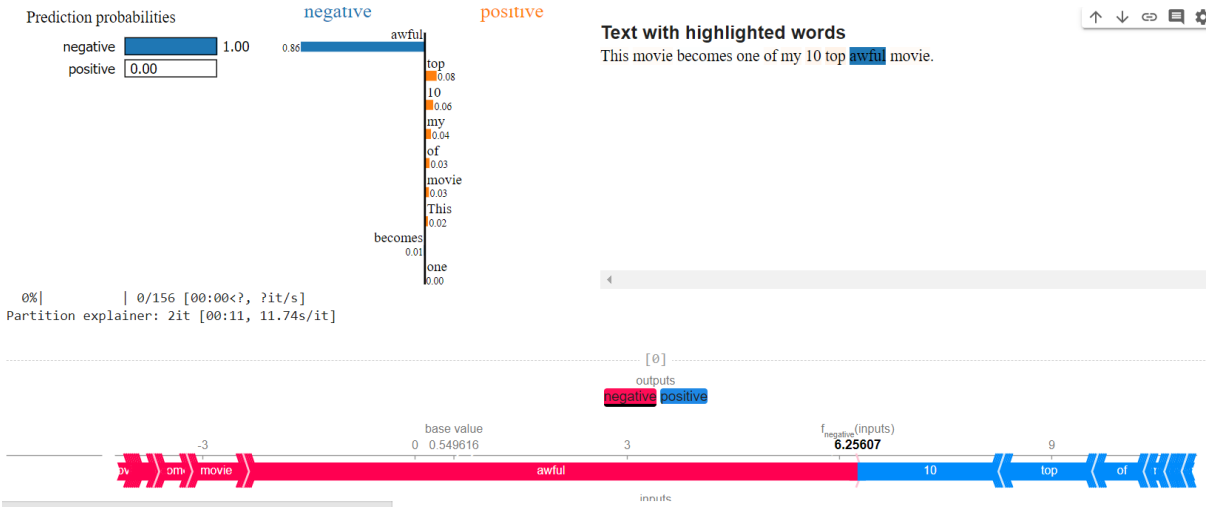



```
<Figure size 640x480 with 0 Axes>
```

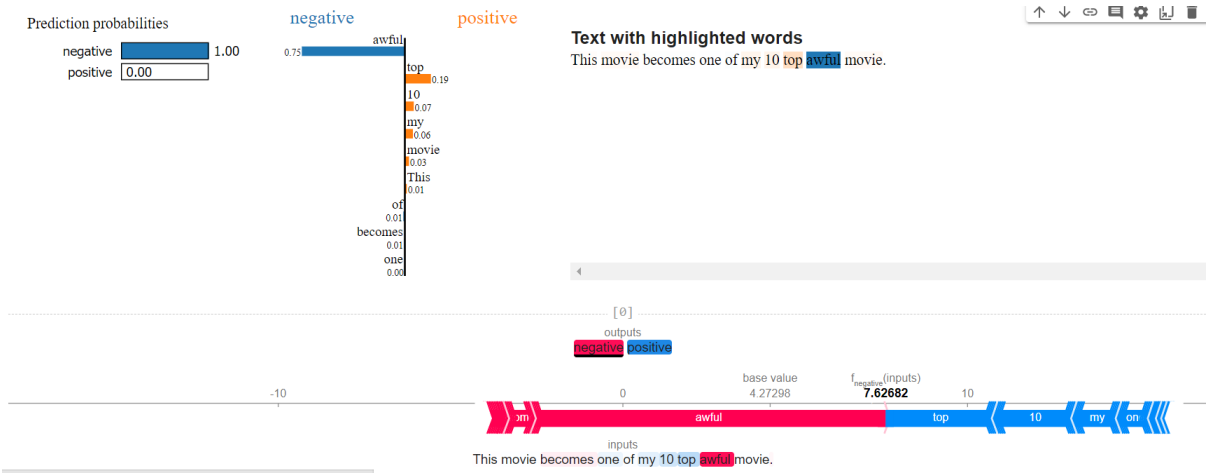
Negative prediction - “This movie becomes one of my 10 top awful movie.”

The comparison between these two models shows similar patterns to the previous observations. However, an interesting aspect is that model 2 calculates the contribution of almost every single word to the prediction.

TA_model_1



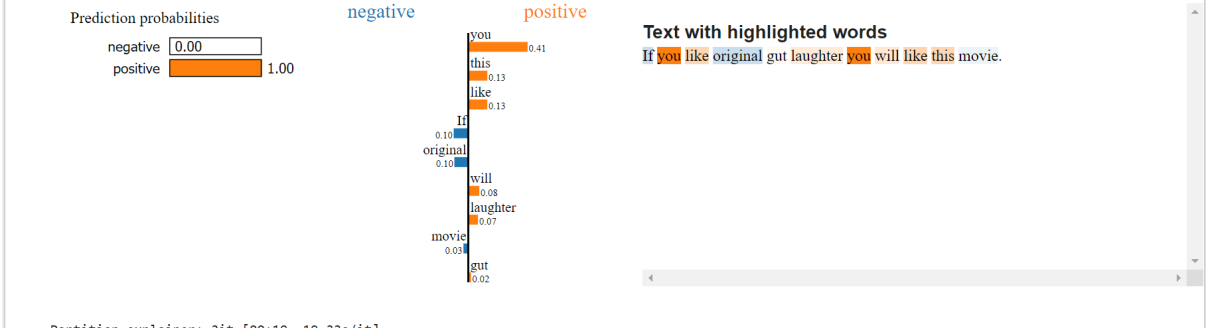
TA_model_2

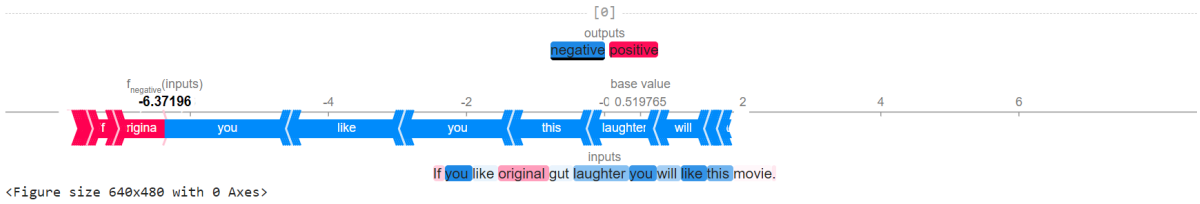


Part 4 NLP Attack

Ex1: "If you like original gut laughter you will like this movie."

```
example1="If you like original gut laughter you will like this movie."
lime(example1)
shapely(example1)
```



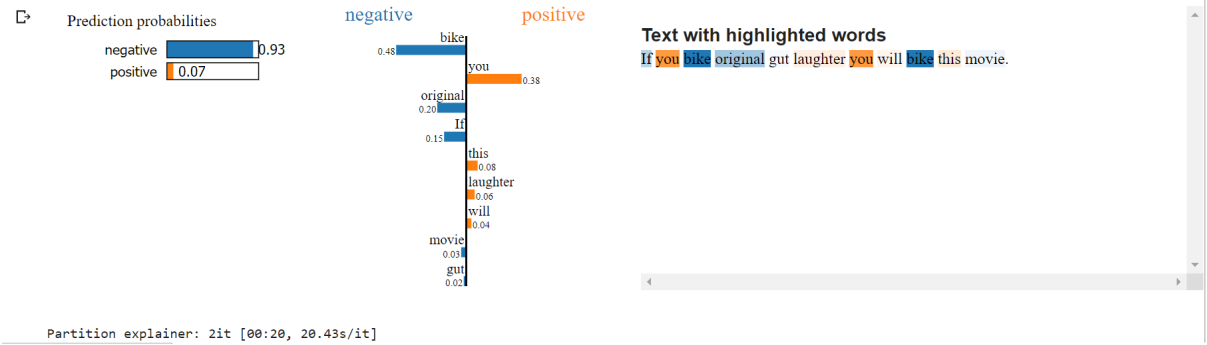



```
<Figure size 640x480 with 0 Axes>
```

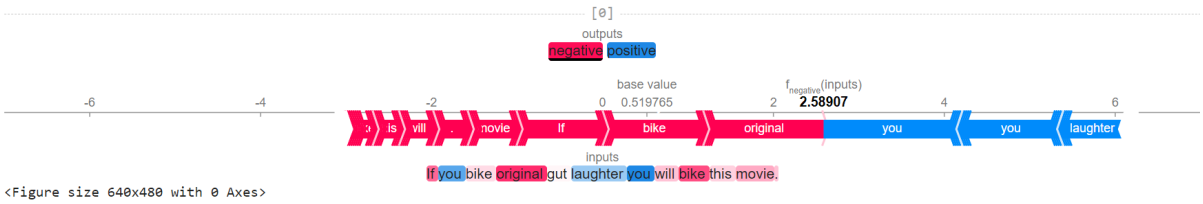
change “like” to “bike”

→ "If you bike original gut laughter you will bike this movie."

```
example1="If you bike original gut laughter you will bike this movie."
lime(example1)
shapely(example1)
```

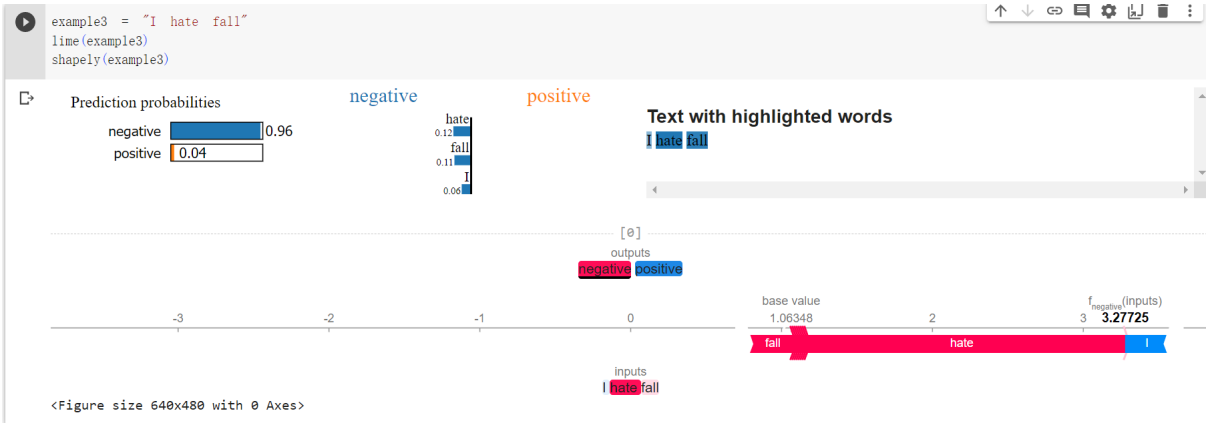



```
Partition explainer: 2it [00:20, 20.43s/it]
```



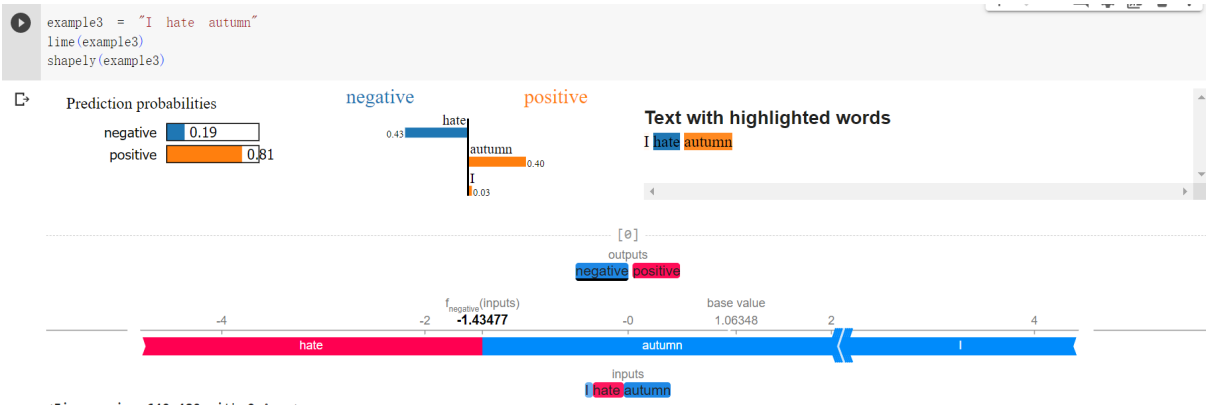
<Figure size 640x480 with 0 Axes>

Ex2,3 : “I hate fall”

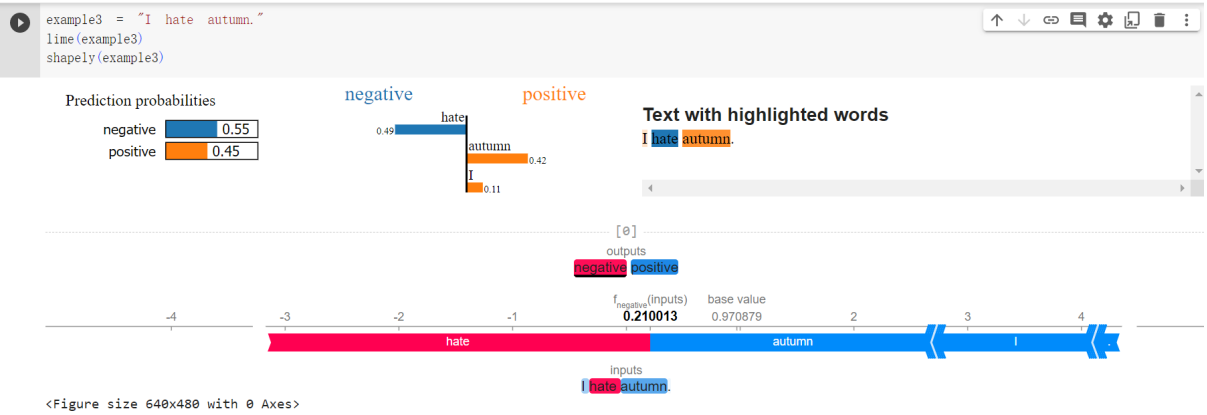


<Figure size 640x480 with 0 Axes>

change “fall” to “autumn” → “I hate autmn”



add. → “I hate autumn.”



Findings

Based on the three examples above, several observations can be made:

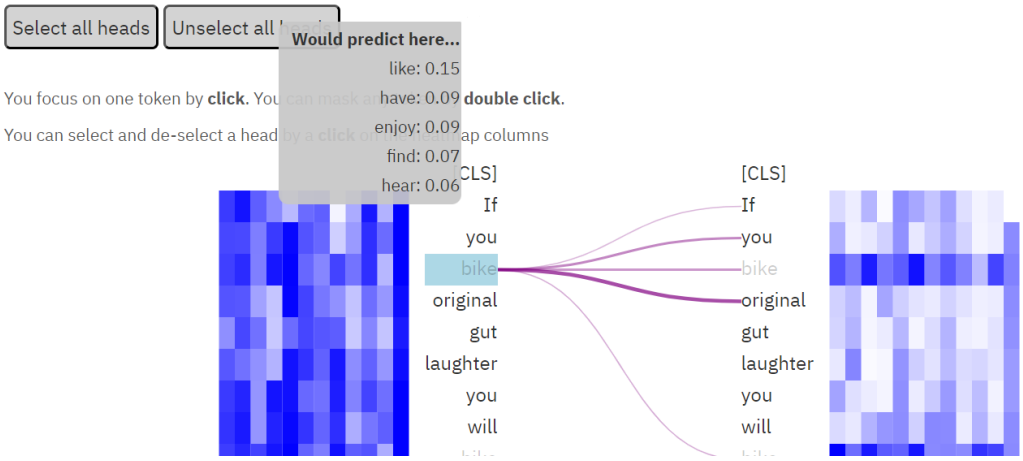
1. Even a single character change can render a sentence unreadable and cause prediction failure.
2. In the case of Ex2, replacing the original word "fall" with its synonym "autumn" resulted in different predictions. Despite the two words being identical and the sentences remaining the same, the sentiment prediction differed. For instance, the sentence "I hate autumn" was predicted as positive.
3. Additionally, in Ex3, the mere addition of a punctuation mark, such as ".", resulted in different predictions.

These observations highlight the sensitivity of language models to subtle changes in text and the potential impact on the predicted outcomes.

Preventions

1. For the Ex1 attack, my approach is to mask each word individually in the sentence and determine whether the masked word appears among the top 5 or 10 predictions. This allows me to identify the type of sentence.

   For example, if I mask the word "bike" but it does not appear among the top 5 predictions, it indicates that my model can detect this type of attack.



2. As for Ex2, my idea to prevent this attack is to have the model learn as many symptoms as possible during its training.