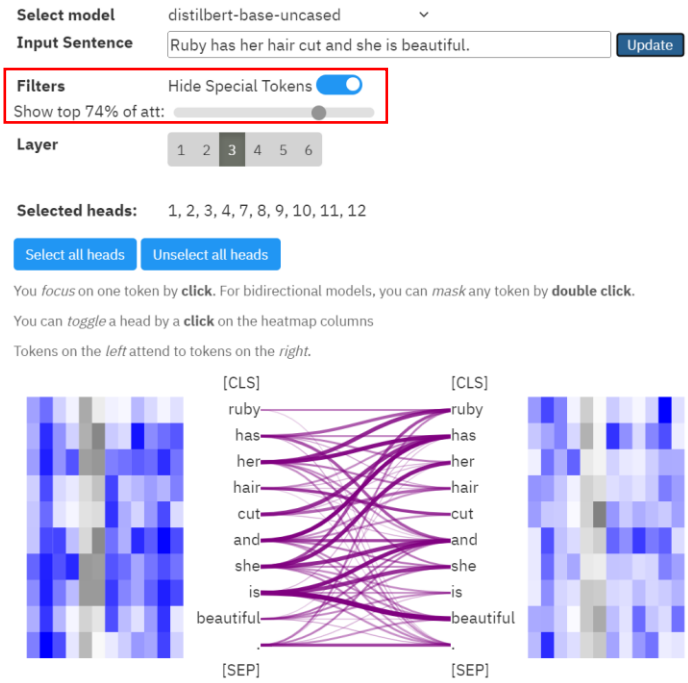
Report of AI HW4　109550027 紀竺均

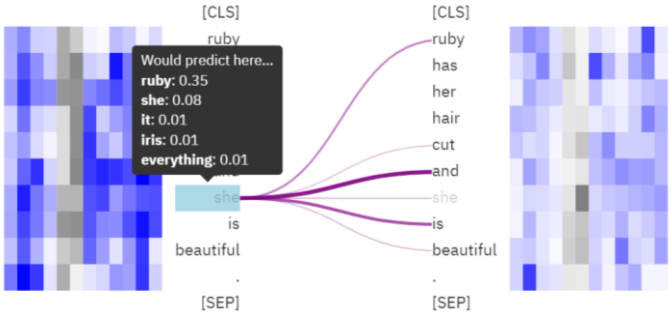
a. Attention Visualization – exBERT

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

exBERT is a visualization tool of Bert that offer concise summaries of info and allow interaction with large models. The attention mechanism helps the model to pay attention on the more important part of the input thus lead to a more precise judgement.



In exBERT, we can select the attention we want to display using the red box and we can display self attention using the blue matrix. Below is a figure of BERT masked-LM training. Here, in the sentence '*Ruby has her hair cut and she is beautiful.*', I masked the word *she* and get some predict words such as *Ruby* and *she* which is both grammarly and semantically correct. Furthermore, the purple lines show that when I masked the word *she,* the model pays attention on words *ruby, cut, and, is, beautiful.* This is an example of attention mechanism.



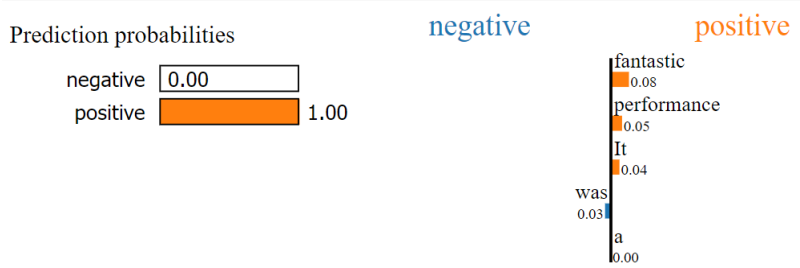
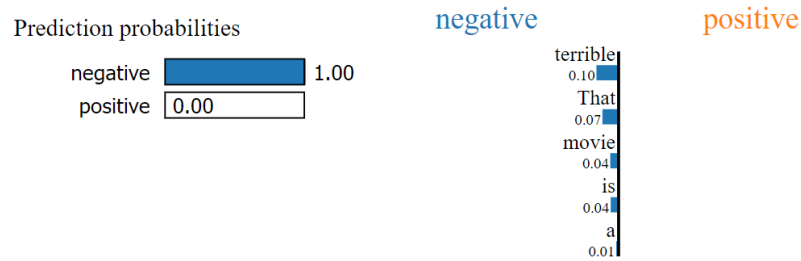
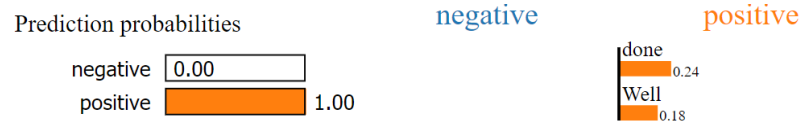
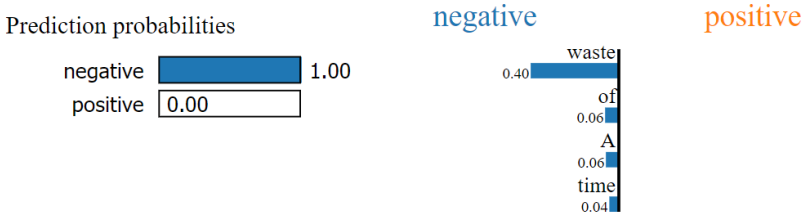
b. Post-hoc Explanation Techniques – LIME

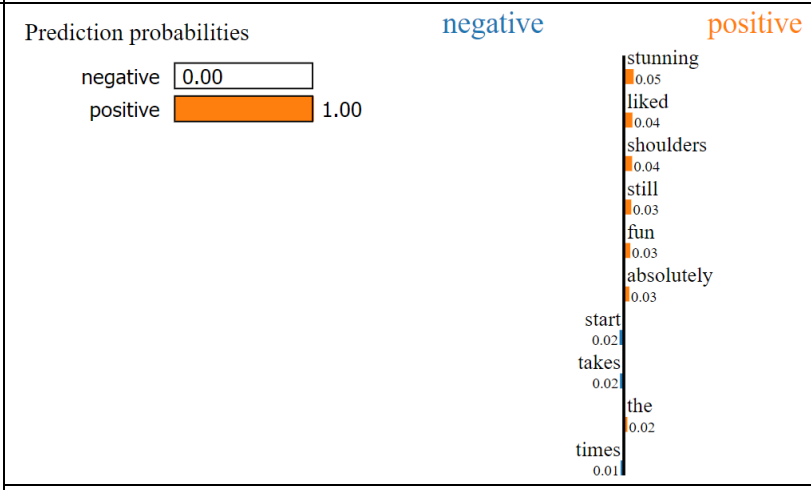
- What is LIME

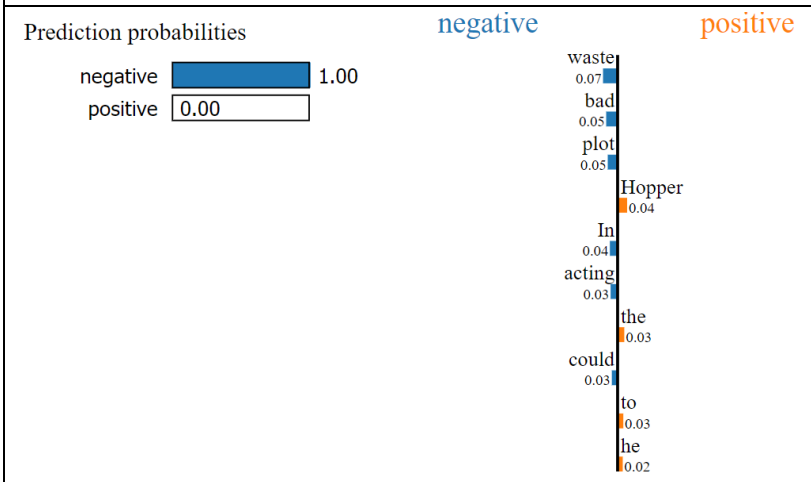
  LIME stands for Local Interpretable Model-Agnostic Explanations, EXPLANATIONS tells us why we can trust this classifier, MODEL-AGNOSTIC tells that LIME treats the model as a black box thus it works for any model, INTERPRETABLE means the explanation is simple for human to understand, finally, LOCAL, relative to global, refers to that we get the explanation locally. LIME can help the users to improve or detect a classifier. Also, it can help us to know more clearly about the models.

- Result of two sentiment classification models

  TA_model_1.pt: distilbert-base-uncased, F1 = 0.93, # dimension = 768

<table>
<tr><td rowspan="3">Sample sentence</td><td>It was a fantastic performance !</td></tr>
<tr><td> 
</td></tr>
<tr><td>That is a terrible movie.</td></tr>
<tr><td> 
</td></tr>
<tr><td rowspan="3">Easy Sentence</td><td>Well done.</td></tr>
<tr><td> 
</td></tr>
<tr><td>A waste of time!</td></tr>
<tr><td> 
</td></tr>
</table>

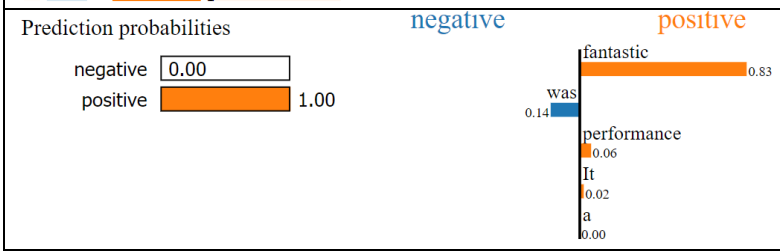
<table>
<tr><td>IMDB example</td><td>What an absolutely stunning movie, if you have 2.5 hrs to kill, watch it, you won't regret it, it's too much fun! Rajnikanth carries the movie on his shoulders and although there isn't anything more other than him, I still liked it. The music by A.R.Rehman takes time to grow on you but after you heard it a few times, you really start liking it.</td></tr>
<tr><td>IMDB example</td><td> 

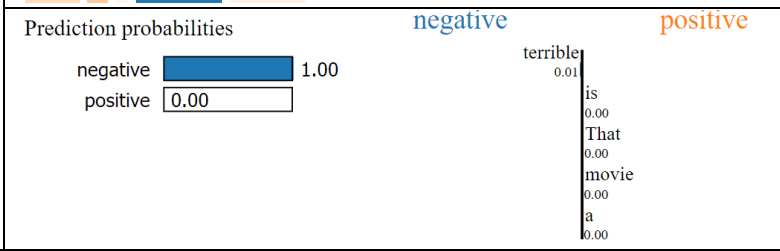
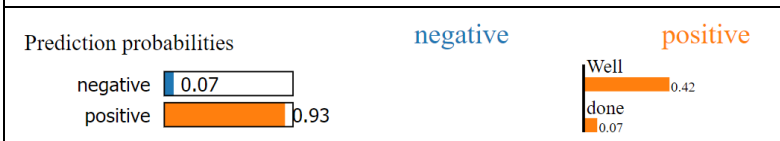
The plot is about the death of little children. Hopper is the one who has to investigate the killings. During the movie it appears that he has some troubles with his daughter. In the end the serial killer get caught. That's it. But before you find out who dunnit, you have to see some terrible acting by all of the actors. It is unbelievable how bad these actors are, including Hopper. I could go on like this but that to much of a waste of my time. Just don't watch the movie. I've warned you.
 
</td></tr>
</table>

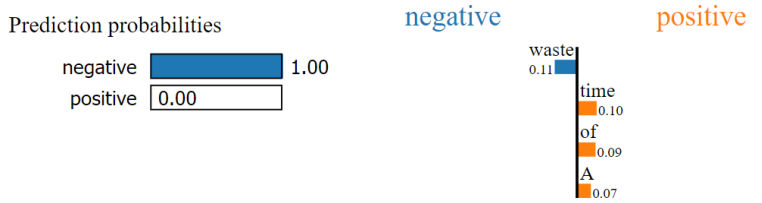
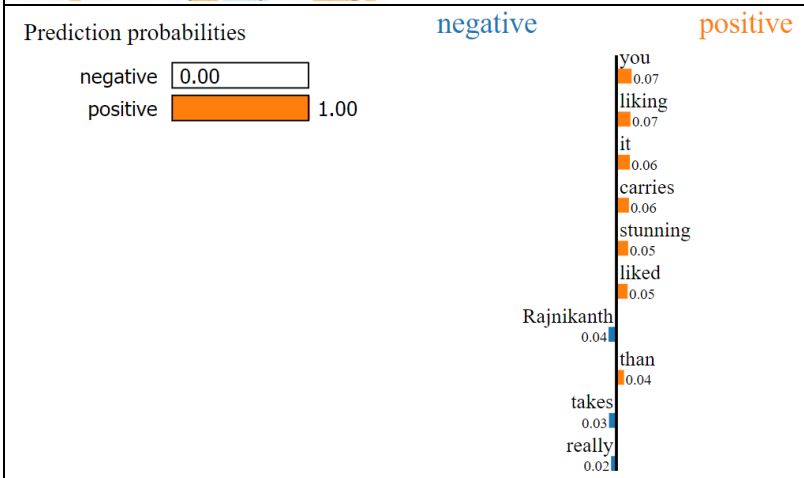
Small conclusion of **distilbert-base-uncased***:*

1. Given the three example (Sample, easy, and IMDB) each with a positive and negative sentence, the result is all perfectly correct.
2. The highlighted words are reliable and meaningful.

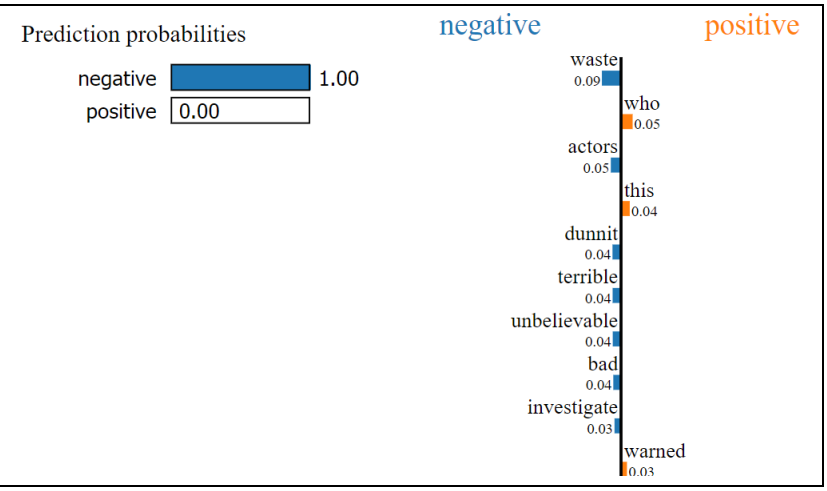
● TA_model_2.pt: prajjwal1/bert-small, F1 = 0.92, # dimension = 512

<table>
<tr><td>Sample sentence</td><td>It was a fantastic performance !
 

That is a terrible movie.
 
</td></tr>
<tr><td>Easy sentence</td><td>Well done.
 

A waste of time!
 
</td></tr>
<tr><td>IMDB example</td><td>What an absolutely stunning movie, if you have 2.5 hrs to kill, watch it, you won't regret it, it's too much fun! Rajnikanth carries the movie on his shoulders and although there isn't anything more other than him, I still liked it. The music by A.R.Rehman takes time to grow on you but after you heard it a few times, you really start liking it.
 

The plot is about the death of little children. Hopper is the one who has to investigate the killings. During the movie it appears that he has some troubles with his daughter. In the end the serial killer get caught. That's it. But before you find out who dunnit, you have to see some terrible acting by all of the actors. It is unbelievable how bad these actors are, including Hopper. I could go on like this but that to much of a waste of my time. Just don't watch the movie. I've warned you.</td></tr>
</table>

IMDB example	

Small conclusion of **prajjwal1/bert-small:**

1. Given the three example (Sample, easy, and IMDB) each with a positive and negative sentence, the result is approximately correct. (‘*Well done*‘ only get 93% positive.)
2. The highlighted words in IMDB example are a little confusing, such as *you* and *Rajnikanth.*

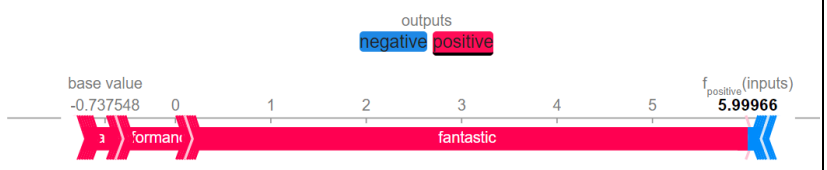
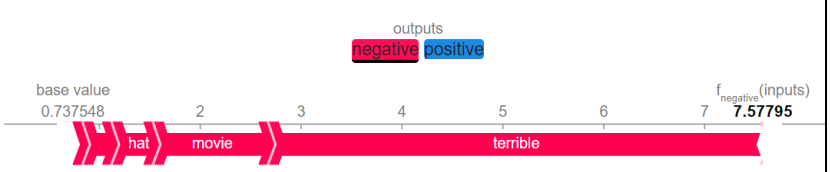
c. Post-hoc Explanation Techniques – SHAP

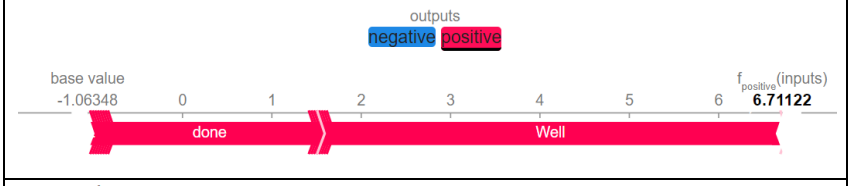
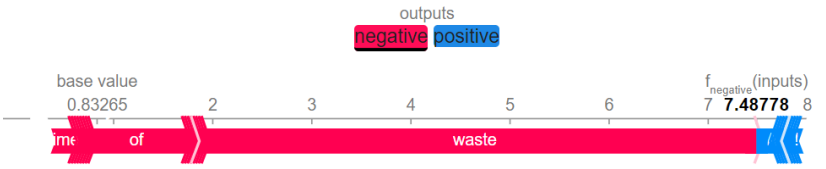
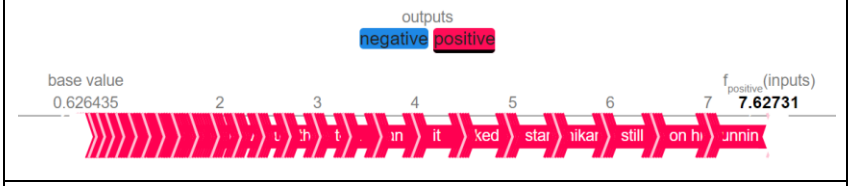
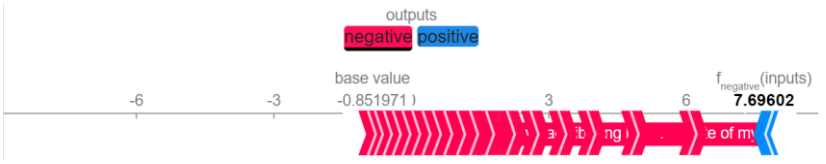
- What is SHAP

  SHAP (Shapley Additive exPlanations) is a mathematical method to explain the ML model. BY calculating the contribution of features (called Shapley values), SHAP can determine the most important features and their influence on the model prediction.

- Result of the two sentiment classification models

  ● TA_model_1.pt: distilbert-base-uncased, F1 = 0.93, # dimension = 768

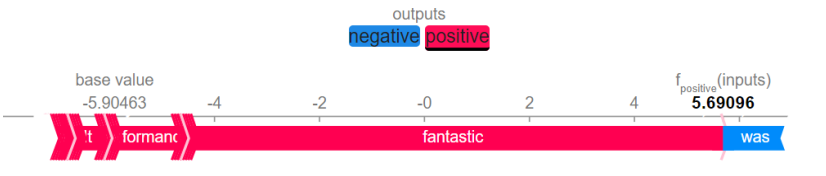
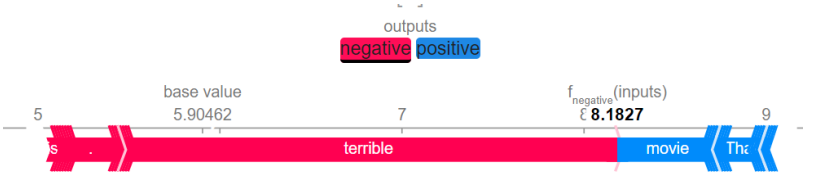
Sample sentence	It was a fantastic performance!  That is a terrible movie. 
Easy	Well done.

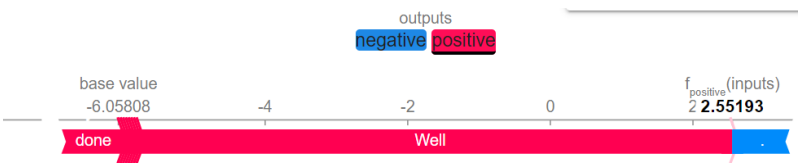
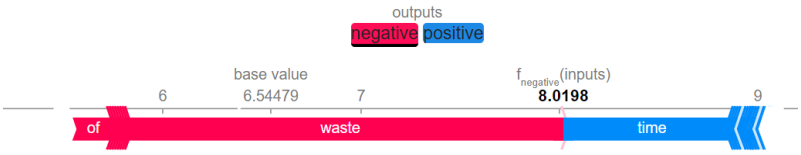
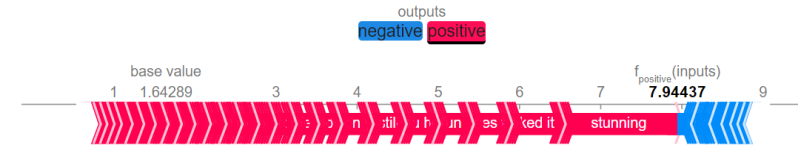
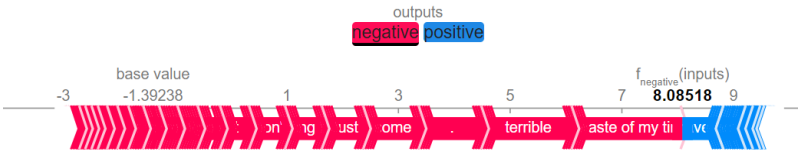
sentence	 A waste of time! 
IMDB example	What an absolutely stunning movie, if you have 2.5 hrs to kill, watch it, you won't regret it, it's too much fun! Rajnikanth carries the movie on his shoulders and although there isn't anything more other than him, I still liked it. The music by A.R.Rehman takes time to grow on you but after you heard it a few times, you really start liking it.  The plot is about the death of little children. Hopper is the one who has to investigate the killings. During the movie it appears that he has some troubles with his daughter. In the end the serial killer get caught. That's it. But before you find out who dunnit, you have to see some terrible acting by all of the actors. It is unbelievable how bad these actors are, including Hopper. I could go on like this but that to much of a waste of my time. Just don't watch the movie. I've warned you. 

Small conclusion of **distilbert-base-uncased***:*

1. The answers are correct. A positive sentence has a higher f-positive value while a negative sentence has a higher f-negative value.
2. In this model, f-value for each sentence is almost greater than six.

● TA_model_2.pt: prajjwal1/bert-small, F1 = 0.92, # dimension = 512

Sample sentence	It was a fantastic performance!  That is a terrible movie. 

<table>
<tr><td rowspan="3">Easy sentence</td><td>Well done.</td></tr>
<tr><td> 
</td></tr>
<tr><td>A waste of time!</td></tr>
<tr><td></td><td> 
</td></tr>
<tr><td rowspan="3">IMDB example</td><td>What an absolutely stunning movie, if you have 2.5 hrs to kill, watch it, you won't regret it, it's too much fun! Rajnikanth carries the movie on his shoulders and although there isn't anything more other than him, I still liked it. The music by A.R.Rehman takes time to grow on you but after you heard it a few times, you really start liking it.</td></tr>
<tr><td> 
</td></tr>
<tr><td>The plot is about the death of little children. Hopper is the one who has to investigate the killings. During the movie it appears that he has some troubles with his daughter. In the end the serial killer get caught. That's it. But before you find out who dunnit, you have to see some terrible acting by all of the actors. It is unbelievable how bad these actors are, including Hopper. I could go on like this but that to much of a waste of my time. Just don't watch the movie. I've warned you.</td></tr>
<tr><td></td><td> 
</td></tr>
</table>

Small conclusion of **prajjwal1/bert-small**:

1. The answers are correct. A positive sentence has a higher f positive value while a negative sentence has a higher f negative value.
2. In this model, f value for each sentence has a relative unstable value, for example, '*Well done.*' only gets 2.5 points.
3. The blue arrow pushes to decrease the prediction, but some blue words seem unreasonable, such as *movie* (pushes to positive), *time* (pushes to positive), *period* (pushes to negative).

d. Conclusion of Comparing two sentiment classification models – **distilbert-base-uncased** and **prajjwal1/bert-small:** (Below I'll call them model 1 and model 2, respectively.)

   From part b and part c, we can observed that using both LIME and SHAP, their explanation is that **distilbert-base-uncased** has better performance than **prajjwal1/bert-small**. My interpretation is due to four main reasons:

   1. The average points in LIME.
      Model 1 > model 2, average point of model 1 = 1.

   2. The highlighted words in LIME.
      Model 1 is more reasonable than model 2.

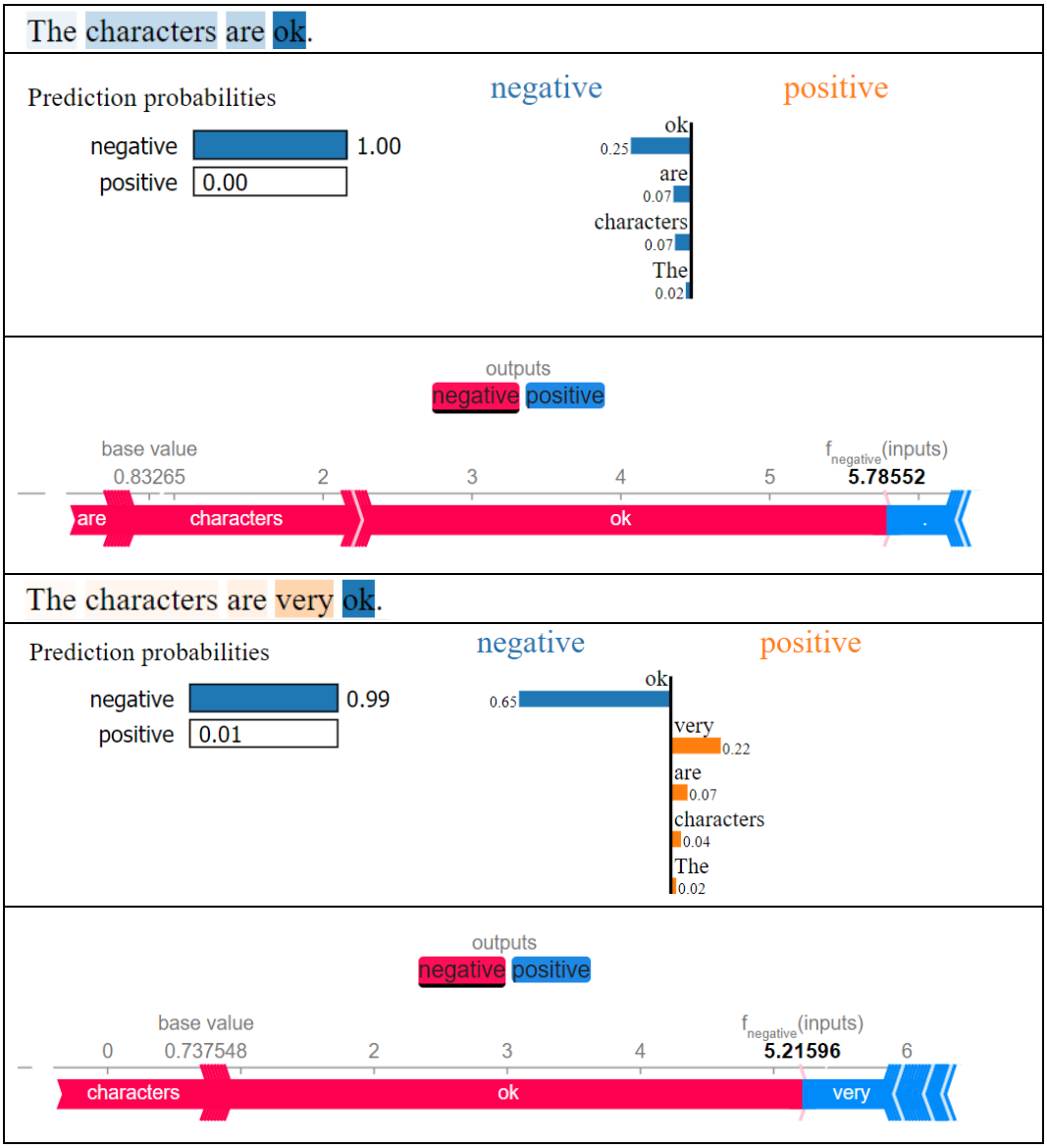
   3. The stability of shapley value in SHAP.
      Model 1 > Model 2. Model 2 has an extreme low value 2.

   4. The blue arrows in SHAP.
      Model 2 has more unreasonable blue arrows than model 1.

e. Compare the explanation of LIME and SHAP.
   From part b and part c, we can easily observed the different appearance of LIME and SHAP. LIME based on words while SHAP highlighted the sentences.
   Other from the appearance, SHAP uses the combination of words to calculate the result value while LIME only calculate a single word's value. SHAP also use arrow-graphs to show the push and pull of the result. In my opinions, BERT relies on the connection of words very much and in this case, SHAP might give a better explanation than LIME.

f. Some input sentences for ATTACK
   From part d, we know that **distilbert-base-uncased** performs better so I will try to attack this model.

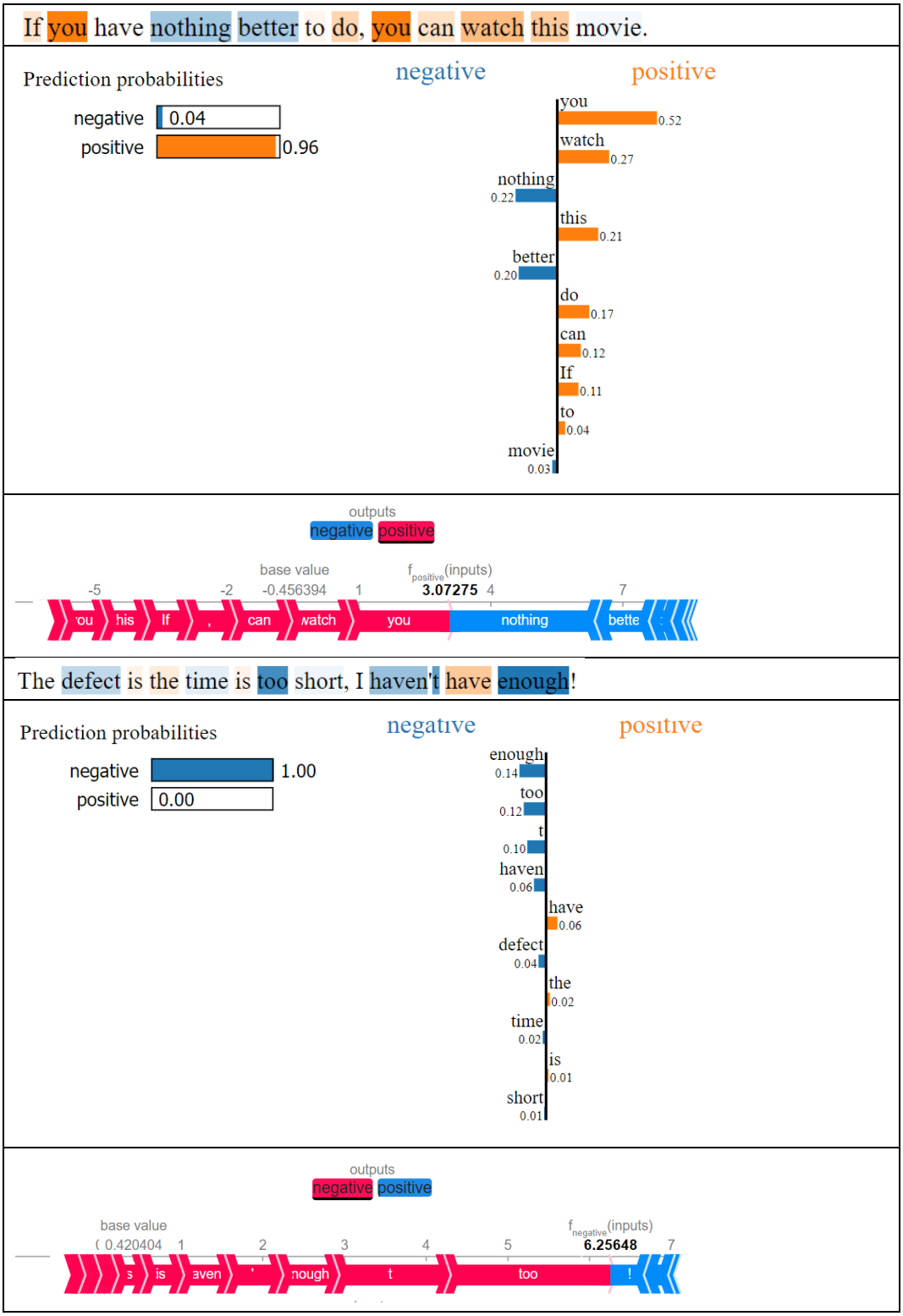
   - First, I try to attack using easily confusing words and sentences. The sentences are:
     1. '*The characters are ok.*'
     2. '*The characters are very ok.*'



Observations:

1. The first sentence '*The characters are ok.*' doesn't seem to give positive or negative sentiment in it, but the prediction is negative. It appears that the model marks '*ok*' as a negative word.
2. So, I give the second sentence '*The characters are very ok.*' to make it a positive sentence. Sadly, the model still predicts it as a negative sentence. From the SHAP graph above, we can observe that though *very* (the blue arrow) tries to push the sentence to the positive side, but the strength is too little. To sum, I consider this attack success.
3. In my opinions, if I train the model next time, I'll tried to reduce this type of error again. That is, tried to connect the word *very* with the word after it. Emphasize the latter if *very* exists. Furthermore, if *very* connects with a word with no sentiment in it, judge the whole sentence to learn the sentiment instead of a single word.

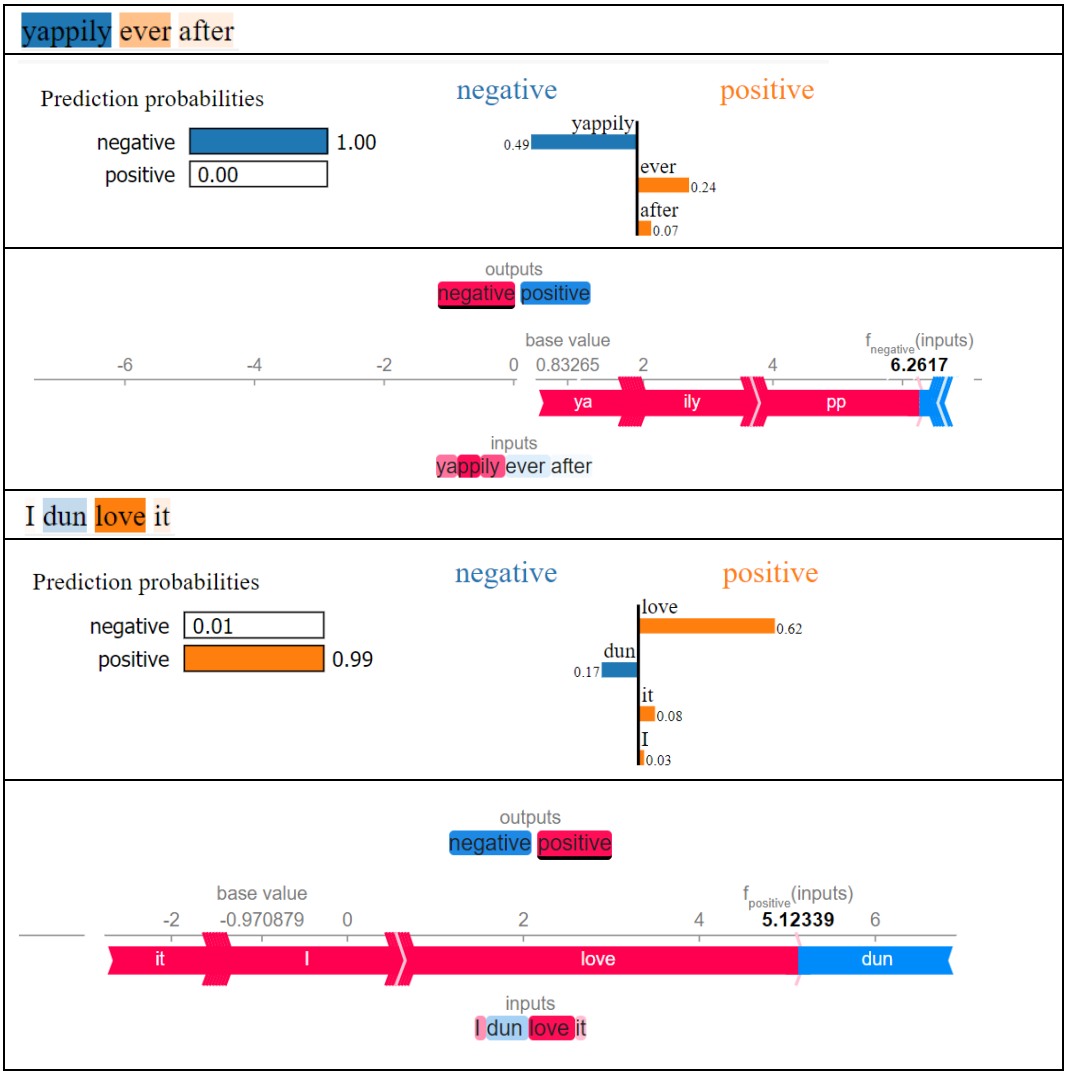
- Next, I tried to attack using the sentences that we have to ‘think’ about it. The sentences are:
  1. *‘ If you have nothing better to do, you can watch this movie.’*
  2. *'The defect is the time is too short, I haven't have enough!'*



Observations:

1. The first sentence ' *If you have nothing better to do, you can watch this movie.*' has some negative meaning. The sentence suggests that there is many things better to do instead of watching the movie. But the model give positive prediction mainly by the sentence '*you can watch this movie*' and didn't think of other possibilities. I consider this attack success.
2. The second sentence '*The defect is the time is too short, I haven't have enough!*' pretends to give negative review, but it actually means that the movie is so good that such time wasn't enough. The model might not think of this step so it gives a negative result. I also consider this attack success.
3. In my opinion, this type of attack is hard to prevent because the art of speaking is way too complicated. But by given much much more training data, things might work out eventually.

- Finally, some typo attack. The sentences are:
  1. mistype '*happily ever after*' into '*yappily ever after*'.
  2. mistype '*I don't love it*' into '*I dun love it*'.



Observations:

1. In both examples, the model give the wrong prediction as I expected. The attacks success.
2. Type error are unavoidable, the model can apply some auto-correct algorithm to avoid typo causes further wrong prediction.

g. Describe problems you meet and how you solve them.

For me, the most difficult part is to think of appropriate attack. I have expected some attack to success but it actually gave the right prediction. I've tried many different methods to make the above attacks success.