Introduction to Artificial Intelligence HW4 Report

110550088 李杰穎

May 27, 2022

1 Attention Mechanism of BERT

Figure 1: Screenshot of ExBERT. [1]

BERT (Bidirectional Encoder Representation for Transformers) is an NLP model based on transformers. BERT is pre-trained on two tasks, masked language model and next sentence prediction.

Masked language model is a task that language model needs to predict the masked word. For example, in following sentence, “The man went to the [MASK] to buy a [MASK] of milk.” There are two words being masked. BERT needs to predict that those two masked words are “store” and “gallon”, respectively.

[1] From original paper `https://arxiv.org/pdf/1910.05276.pdf`

Next sentence prediction is a task that model will be given two sentences, let's say sentence A and B. It needs to tell us is B the next sentence of A.

In this section, I will use ExBERT, a visualizing tool for different variances of BERT, including `bert-base-cased` and `distilbert-based-uncased`, to understand the attention mechanisms of BERT.

1.1 Using BERT as Masked Language Model

Masked Language Model (MLM) is a task that its input is a sentence with part of words being masked. The goal of language model is to predict the masked words using the context of sentence.

In this section, I will use $\text{BERT}_{\text{base}}$ as language model.

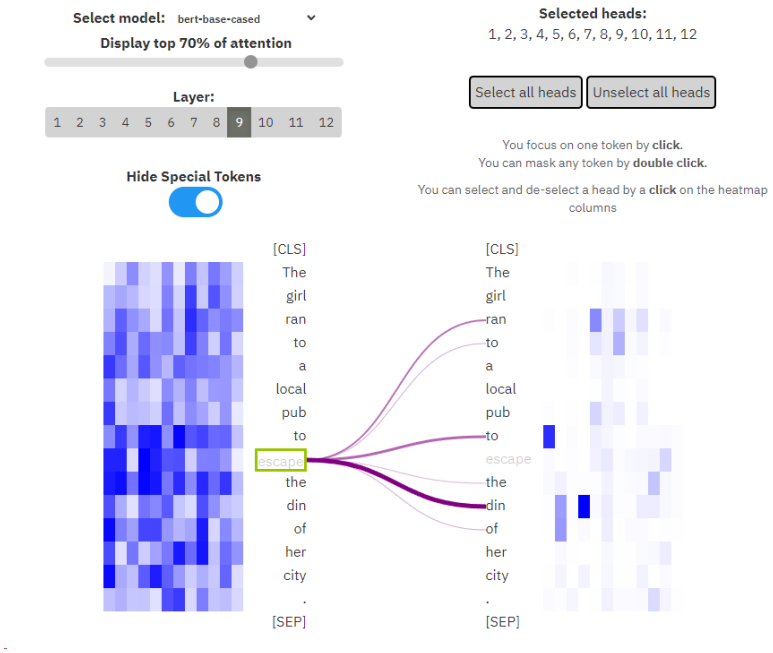


Figure 2: Using the sentence "The girl ran to a local pub to escape the din of her city." and mask the word "escape".

As we can see in Figure 2, if we mask the word "escape" and let BERT predict which word should appear here. In layer 9 of BERT, the words that get attention are "ran", "to" and "din". These words are exactly the words that is relevance with the masked word "escape". This is an

example of attention mechanism.

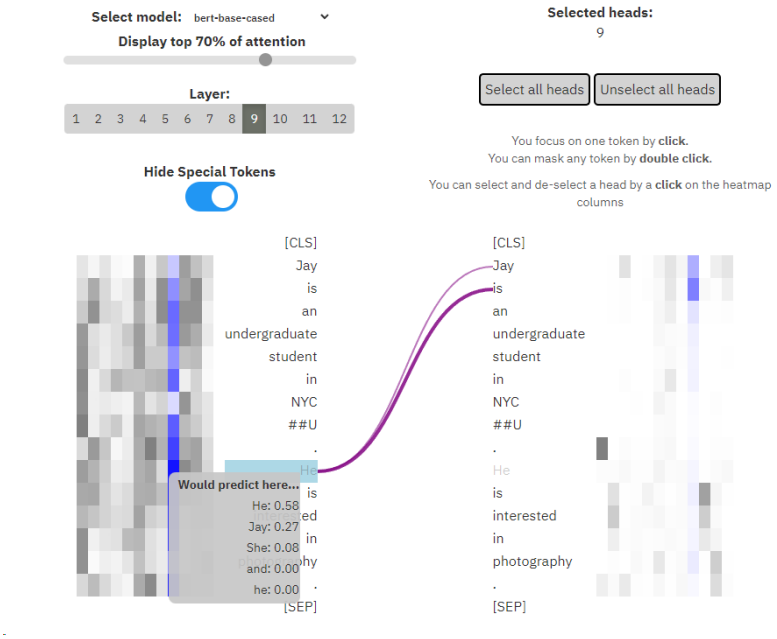


Figure 3: Using the sentence "Jay is an undergraduate student in NYCU. He is interested in photography." and mask the word "He".

Another example is shown as Figure 3. I mask the subject and see how BERT know what word should be filled in. As a normal human would do. BERT pay its attention on the subject of the first sentence, which is my name "Jay". We can also observe that "He" has the highest chance to appear in the masked position. This also shows that BERT knows Jay is usually the name of a male.

The last example is to show that BERT is able to know that who does a specific pronoun refer to. As we can see in the Figure 4, BERT pays its attention at "Mary" when it predict the masked word, instead of "Jay". This shows that BERT can analysis the grammar structure in the sentence and know which word should be paid attention at.

In conclusion, the examples above show that how attention mechanism work in BERT. And indeed, this mechanism helps BERT perform better compared with other methods like n-gram model or ELMo.

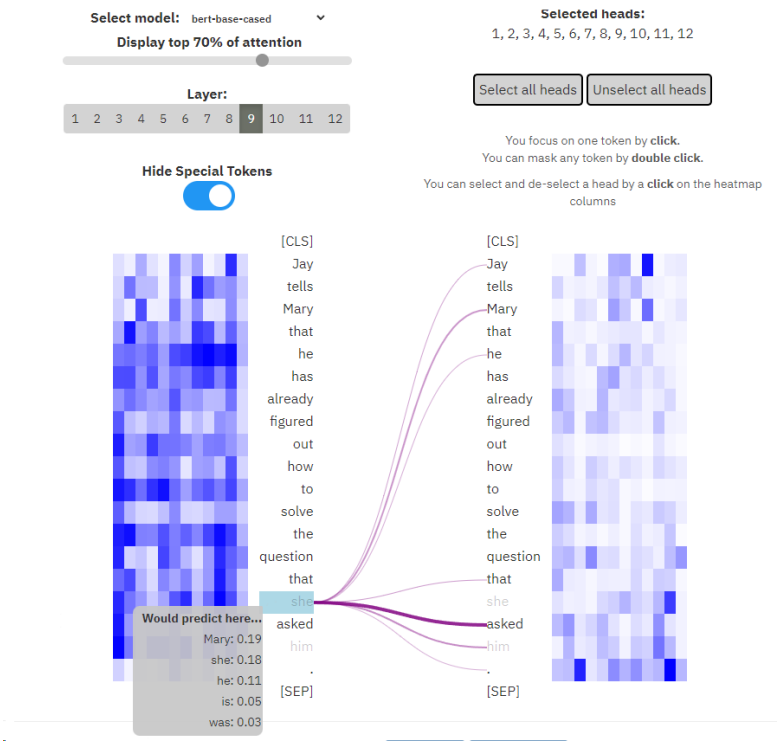


Figure 4: Using the sentence "Jay tells Mary that he has already figured out how to solve the question that she asked him." and mask "she" and "him" in the second sentence.

2 Comparison of BERT and DistilBERT

DistilBERT is smaller version of BERT. The basic idea of DistilBERT is to use a light-weight model architecture to "mimic" the behavior of the original BERT. The author of DistilBERT claimed that this method can "reduce the size of a BERT by 40%, while retaining 97% of its language understanding capabilities and being 60% faster."[2]

In ExBERT, we can use both `bert-base-cased` and `distilbert-base-uncased`. Thus, we will also use ExBERT to compare these two model.

To compare DistilBERT with BERT. I use the same three sentences as in previous section and mask same words. Below are the experiments results. Noticed that these results are all from layer 3 of DistilBERT. This is because after some observations, I think this is layer that DistilBERT pay attention to the context of sentence. Other layers are paying attention to either the previous word or the next word or the [CLS] tag.

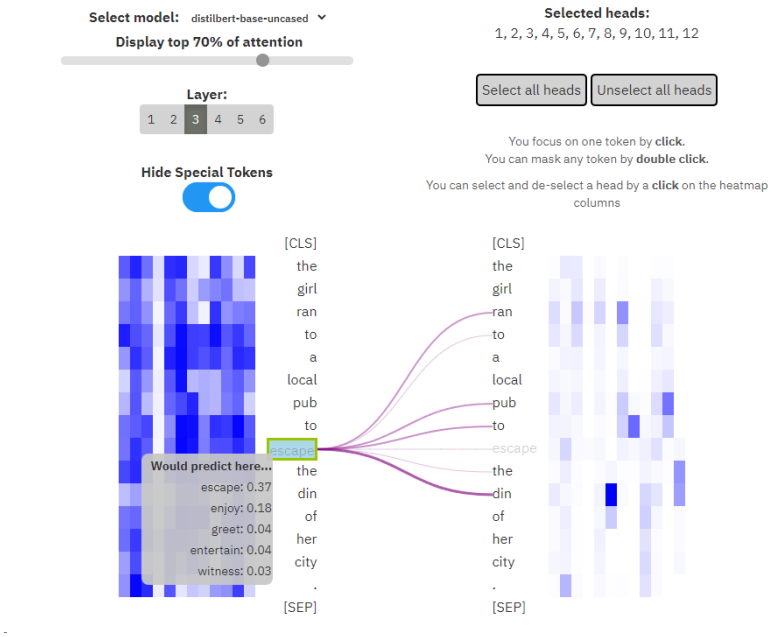


Figure 5: Using the sentence "The girl ran to a local pub to escape the din of her city." and mask the word "escape".

As we can see in Figure 5, DistilBERT is able to predict correctly, and the words it pays attention to is similar to those BERT pays attention to. This shows that DistilBERT learn some "knowledges" from the original BERT.

[2]From abstract of original paper. `https://arxiv.org/abs/1910.01108`

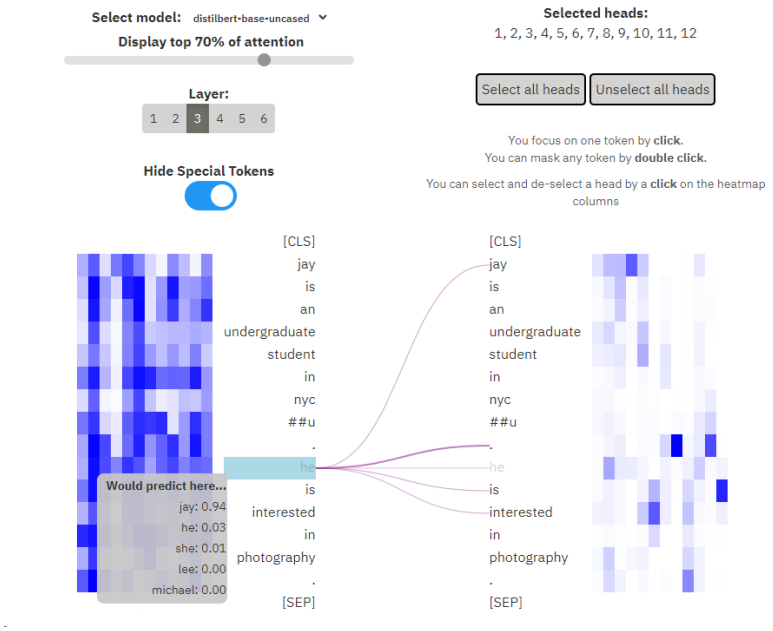
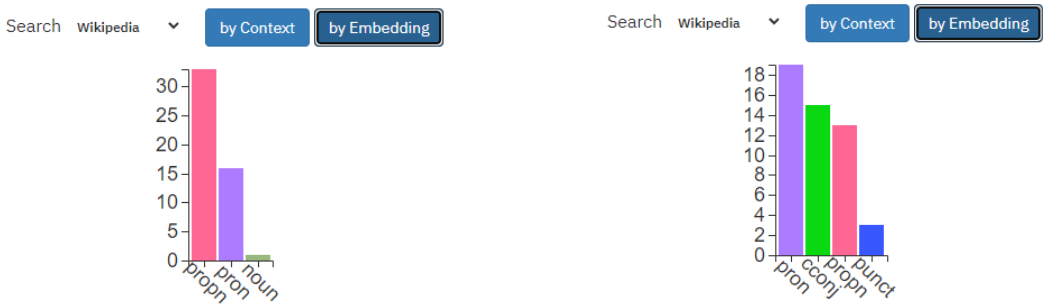


Figure 6: Using the sentence "Jay is an undergraduate student in NYCU. He is interested in photography." and mask the word "He".



(a) DistilBERT (b) BERT

Figure 7: The matched word summary of the sentence "Jay is an undergraduate student in NYCU. He is interested in photography." and mask the word "He".

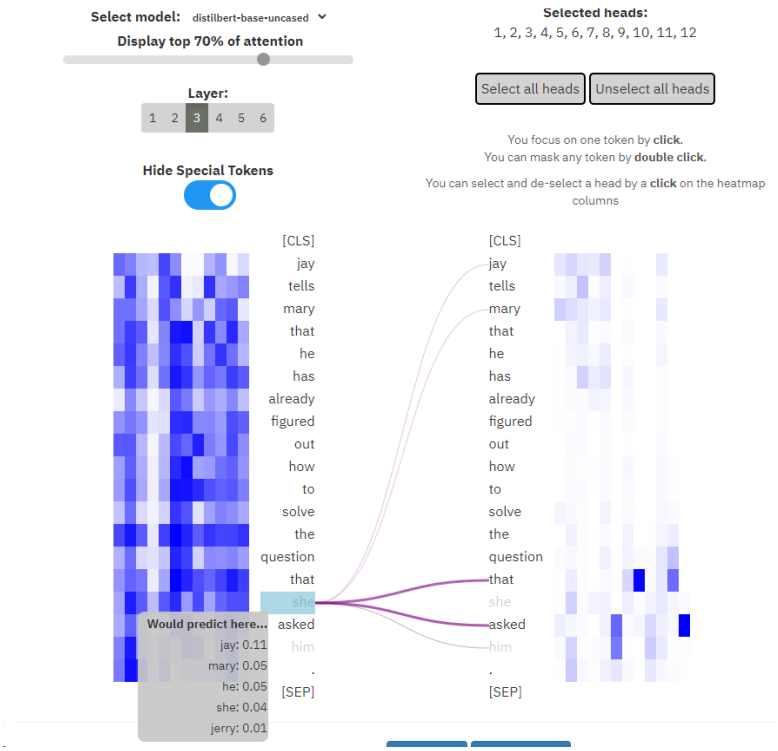


Figure 8: Using the sentence "Jay tells Mary that he has already figured out how to solve the question that she asked him." and mask "she" and "him" in the second sentence.

But if we look at more complicate examples, like sentences in Figure 6 and Figure 8, DistilBERT can't predict the masked word as well as original BERT can do.

Let's first look at Figure 6. DistilBERT can predict that the masked word can be the name "Jay" or "He". Although both of these is correct answer and model knows that Jay is a male name. But the better answer is "He", which is closer to daily English usage. But DistilBERT only has 3% of probability to fill "He" here. This shows that DistilBERT perform worse than original BERT.

Furthermore, we can also use the "by Embedding" button to see the matched word. Figure 7a shows that the matched word summary in the last layer of DistilBERT. In contrast, Figure 7b shows that the matched word summary in the last layer of BERT. We can see that BERT matched more pronoun than DistilBERT does, which is a better matched compared with proper noun. This also shows the difference between BERT and DistilBERT.

Figure 8 is the last example. Compared with the result shown in Figure 4, DistilBERT's prediction is completely wrong, the most possible word is "Jay", which is the last word to be filled in. This example shows that there has a significant performance difference between BERT and DistilBERT.

In conclusion, although DistilBERT can predict words correctly in some simple sentences, but if the sentences are too complicate, DistilBERT may not perform very well.

3 Explainable AI: LIME and SHAP

Nowadays, people have trained lots of AI models to do various of tasks. But most of these models are black boxes, we don't know how machine makes its decision. This makes us hard to trust the prediction of machine. Even worse, if the models make some wrong decision, it's nearly impossible to debug those AI models. Therefore, we need to conceive methods to explain behavior of AI models. In this section, I will discuss two methods that can explain AI models. One of them is local interpretable model-agnostic explanations (LIME) and another is shapley additive explanations (SHAP).

3.1 Local Interpretable Model-agnostic Explanations (LIME)

LIME is an method to explain AI model. Its name clearly shows its properties.

1. **Local**: LIME explain the given prediction locally.
2. **Interpretable**: The explanation given by LIME is easy enough for human to understand.
3. **Model-agnostic**: Treat the model as a black box so that LIME works on every model.

Given a prediction we want to explain, LIME first creates a datasets by perturbed the given input. Then it uses this datasets to train a simple (often a linear classifier) and explainable classifier. The performance of this classifier need to be good enough on the new datasets. In this way, we can achieved the so-called "local fidelity". Figure 9 illustrates above procedures. $g(z')$ is the classifier we want to train.

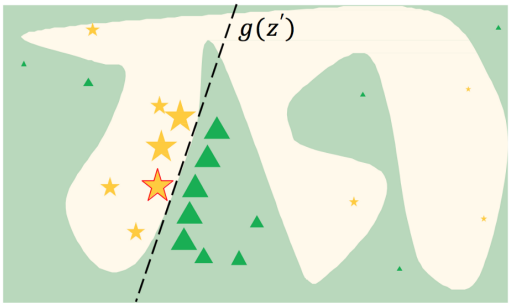


Figure 9: A simple classifier.[3]

We can noticed that LIME only works on a single prediction. We can't get a global explanation using LIME. On the contrast, the next method I will introduce works both on local and global explanation.

3.2 Shapley Additive Explanations (SHAP)

SHAP is an method based on Shapley value. Shapley value is first proposed by an American mathematician Lloyd Stowell Shapley. Shapley value can be used to compute the contribution to the final output of model for a given feature. The following equation is for computing the Shapley value for a given feature j.

[3]From `https://reurl.cc/3o4YE8`

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}} \left(x_{S \cup \{i\}} \right) - f_S \left(x_S \right) \right] \tag{1}$$

Where F is all the features used to build the model. $f_S(x_S)$ is the output of the model trained with feature set S.[4]

We can also noticed that SHAP deals with features, it can also give the global explanation.

3.3 Explanations Given by LIME and SHAP

Before using the sentences from IMDB Datasets to test LIME and SHAP. Let's first look at some simpler sentences. And also, I will test two models given by TAs, which are `distilbert-base-uncased` and `prajjwal1/bert-small`. And because these two models are trained on IMDB datasets to classify the sentiment of given movie review. Therefore, the sentences that I will test in experiments all express strong sentiments.

3.3.1 Using Simple Sentences

I will use four sentences

1. It was a fantastic performance!
2. That is a terrible movie.
3. This movie is a must-see.
4. This movie really brings art to a new level.

Below is the explanation of four sentences given by LIME and SHAP. Model used here is `distilbert-base-uncased`.

As we can see from Figure 10 to Figure 17, both LIME and SHAP can explain the prediction very well. Both method can give us the contribution of each word to the result. Even a harder

[4]For detailed explanation, please refer to the original paper or this article. `https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html`

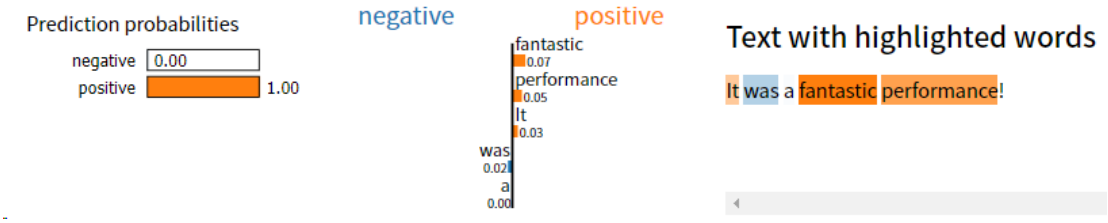


Figure 10: The explanation of first sentence given by LIME.

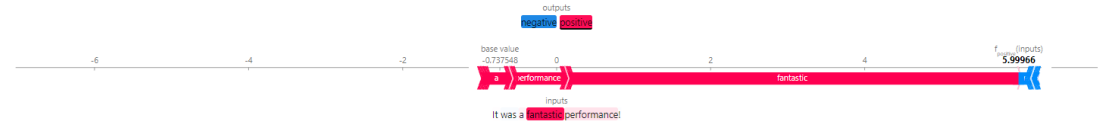


Figure 11: The explanation of first sentence given by SHAP.

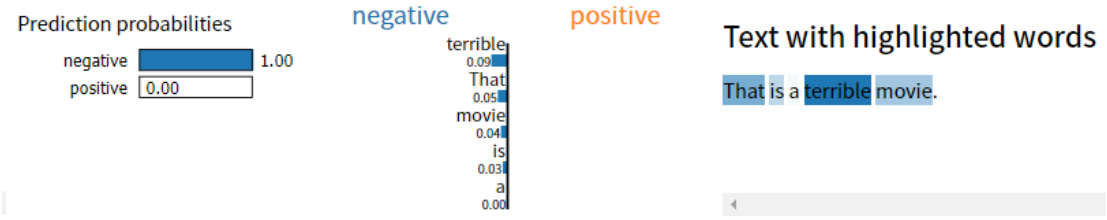


Figure 12: The explanation of second sentence given by LIME.

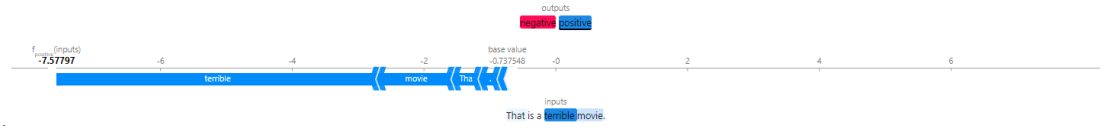


Figure 13: The explanation of second sentence given by SHAP.

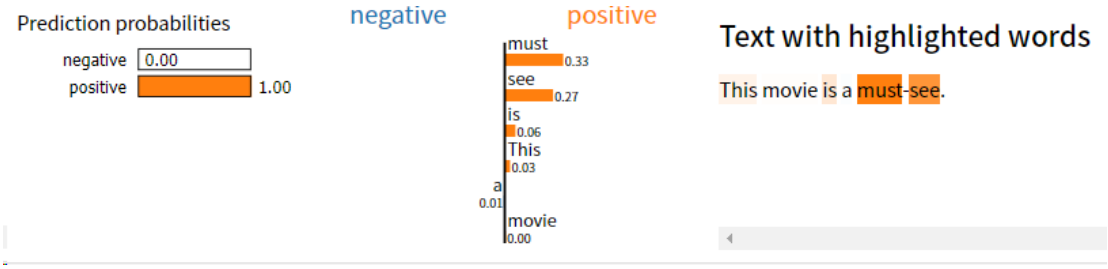


Figure 14: The explanation of third sentence given by LIME.

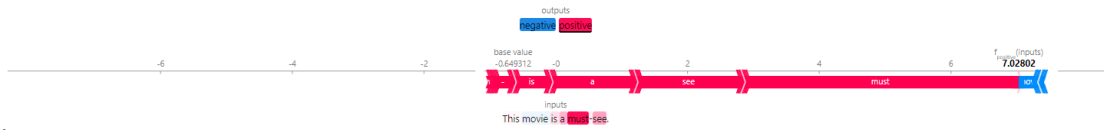


Figure 15: The explanation of third sentence given by SHAP.

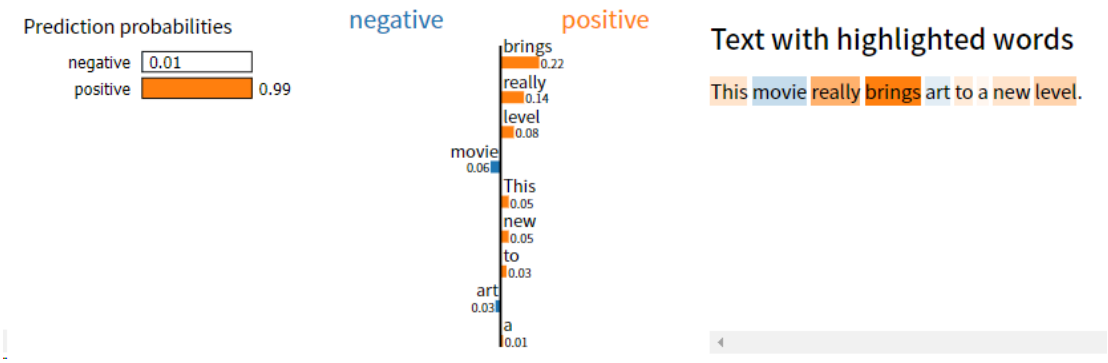


Figure 16: The explanation of fourth sentence given by LIME.

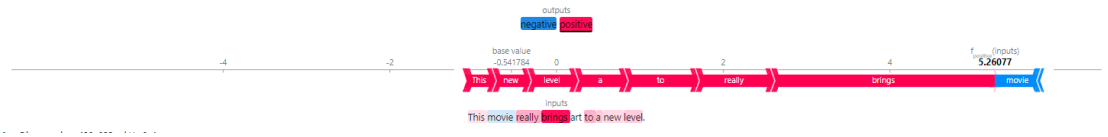


Figure 17: The explanation of fourth sentence given by SHAP.

sentence like the third and fourth sentences (Figure 14 to Figure 17), which doesn't mention any sentimental word, both methods can know that "brings art to a new level" and "must-see" has positive meaning and is important to the final prediction.

3.3.2 Using Sentences from IMDB Datasets

In this section, I will use the reviews in test set of IMDB Datasets. I've picked four reviews, I will list them below. The first and second reviews are positive samples and rest of them are negative samples.

1. I bought the DVD a long time ago and finally got around to watching it.I really enjoyed watching this film as you don't get the chance to see many of the more serious better quality bollywood films like this. Very well done and but I would say you need to pay attention to what is going on as it is easy to get lost. When you start watching the movie, don't do anything else! I would actually advise people to read all the reviews here...including the ones with spoilers, before watching the movie. Raima Sen gave her first great performance that I have seen. Aishwarya was easily at her best. All performances were strong, directing and cinematography...go watch it!

2. A ghost story on a typical Turkish high school with nice sound and visual effects. Taylan biraderler(taylan brothers) had made their first shots on a tv-show a couple of years ago, as far as i know. That was kind of a Turkish X-Files, they had very nice stories but lacked on visual effects. This time it seems they had what they needed and used them well. This movie will make you laugh that's for sure, and as well it might have you scared. It has a nice plot and some young, bright actors in it. If you are a high school

student in Turkey you will find so many things about you here. There are many clues in the movie about its story and ending, some you understand at the moment, some will make sense afterwords, the dialogs were written very cleverly. So these make the movie one of the best Turk movies made in the last years. Do not forget, this movie is the first of its kind in the Turkish film industry.

3. This movie is so unreal. French movies like these are just waste of time. Why watch this movie? Even, I did not know..why. What? The well known sex scene of half-siblings? Although the sex scene is so real and explicit, but the story it is based upon is so unreal. What is the use of it, then? Can you find easily in life, half sibling doing such things?

Did I learn something from this movie? Yeah: some people are just so fond of wasting time making such movies, such stories, such non-sense. But for those who like nihilism, nothingness in life, or simply a life without hope, then there you are.. you've got to see this movie.

Only one worth adoring, though: CATHERINE DENEUVE. She's such a strikingly beautiful woman.

4. I saw this with high expectations. Come on, it is Akshay Kumar, Govinda, and Paresh Rawal, who are all amazing at their comedy, I was really hoping for a laugh riot. Sadly, that is not what I got at all...

Unfortunately, nothing in this movie really made me laugh out loud. There were times when I chuckled at one or two things, but nothing really made me laugh. In short, it was badly attempted comedy, and in a way, a bit of a Hera Pheri wannabe.

Out of the three main guys, I think Paresh Rawal's role was the most powerful. It wasn't the biggest role, but it certainly stood out more than Govinda or Akshay. Their performances were okay I guess. Nothing special, just mediocre. Though Govinda stole the limelight from Akshay in more than a few scenes. Lara Dutta and Tanushree Dutta also make appearances in this film, and both of them were pretty bad. Lara's role did not move me, or make me laugh, and Tanushree Dutta's character just got on my nerves! The music seems to be the only good thing about Bhagam Bhag. My favourite song is "Tere Bin", followed by "Afreen", which I really liked. "Signal" and the title song "Bhagam Bhag" are also worth a listen.

You either will like it or you won't. And judging by the poor comedy and lack of direction, I don't think you will.

Below is the explanation of four reviews given by LIME and SHAP. Model used here is also `distilbert-base-uncased`.

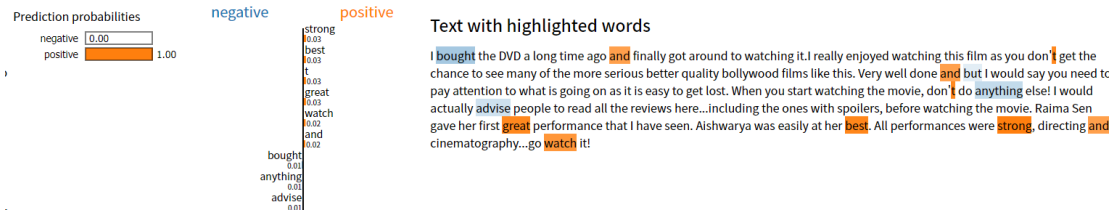


Figure 18: The explanation of first review given by LIME.

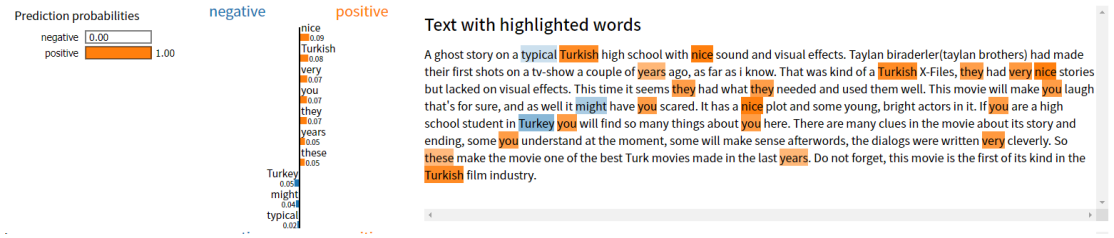


Figure 19: The explanation of second review given by LIME.

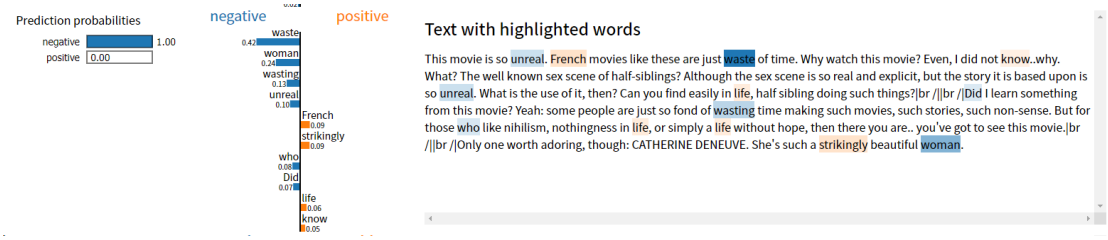


Figure 20: The explanation of third review given by LIME.

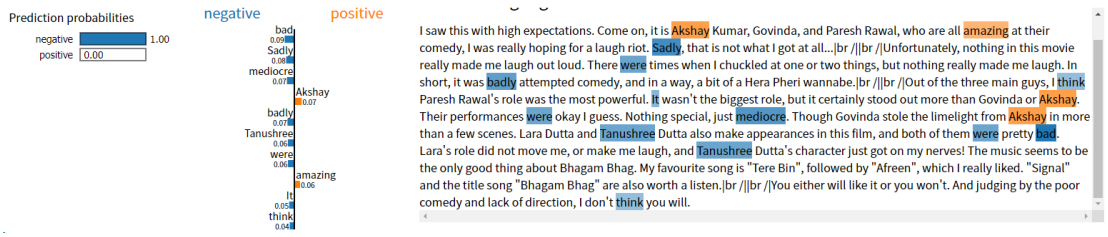


Figure 21: The explanation of fourth review given by LIME.

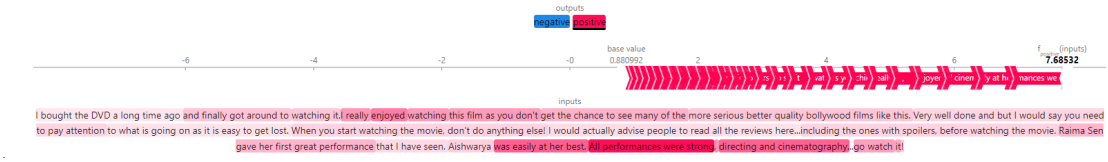


Figure 22: The explanation of first review given by SHAP.

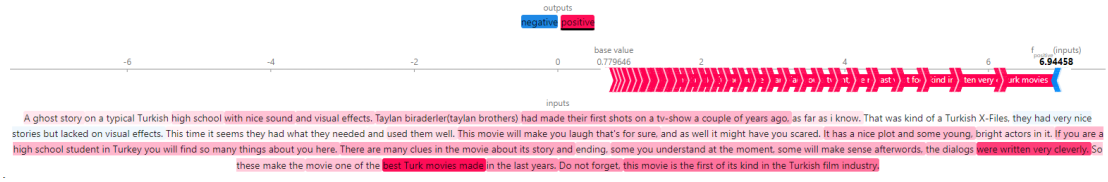


Figure 23: The explanation of second review given by SHAP.

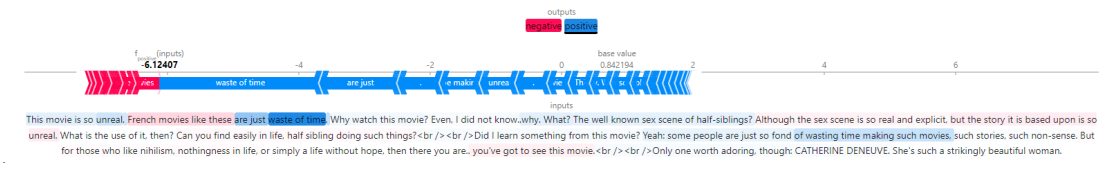


Figure 24: The explanation of third review given by SHAP.

Figure 25: The explanation of fourth review given by SHAP.

As we can observe in Figure 18 to Figure 25. Although LIME and SHAP both can explain short sentence really well. However, when it comes to the explanation of long sentences like movie reviews in IMDB datasets, explanation given by LIME are worse than those given by SHAP. LIME can only marked some single words instead of whole sentence like SHAP give us. To be worse, some words that LIME marked are unreasonable. For example, like in Figure 20, LIME marked the last word "woman", and said that it's contribution to the negative prediction is second high among all the words, which is absolutely wrong. On the contrary, Figure 24 shows the explanation of third review given by SHAP. As we can see, this explanation is way better than LIME gives. It marks the word by phrases and those which get high Shapley value are indeed the main reason why this review is negative. It can also mark the positive part of a negative review. Therefore, I think SHAP is a better explanation techniques than LIME when it comes to sentimental classification.

And I think the reason why LIME's explanation of long sentences is worse than short sentence is mainly because LIME treats every word in the review as a feature. Therefore, if there are too many features (like in movie review), it's very hard to generate a stable perturbed datasets. And also, we can't train a simple classifier to approximate the original model.

Another disadvantage of LIME is that LIME generates the perturbed datasets randomly. Therefore, the explanation given by LIME is different. Figure 26 shows this effect. I input same review in Figure 18, but the final explanation is completely different.

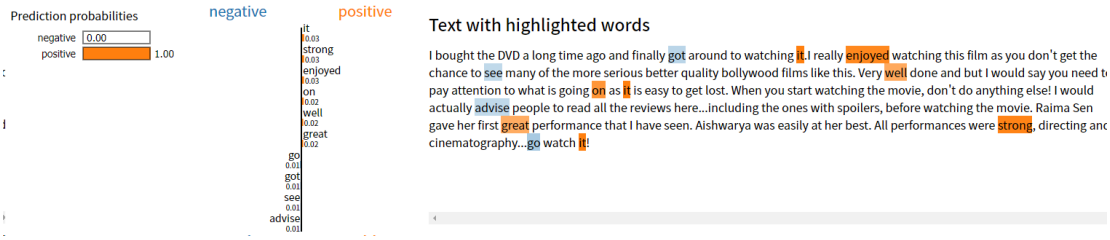


Figure 26: The randomness in explanation give by LIME.

In conclusion, I personally think that SHAP is a better explanation technique than LIME.

4 Attack NLP Model

In this section, I will try to attack the model that TAs give. I will mainly use the first model which is `distilbert-base-uncased`.

4.1 Testing sentences that doesn't have sentiment in it

Let's first look at a simple example, "This is an ordinary movie.". This example doesn't have obvious sentiment in it. But as we can see in Figure 27, model predicts that it's negative. I don't know it's because the IMDB datasets doesn't have these examples that doesn't have obvious sentiment in it or it's model think "ordinary" is a negative word.

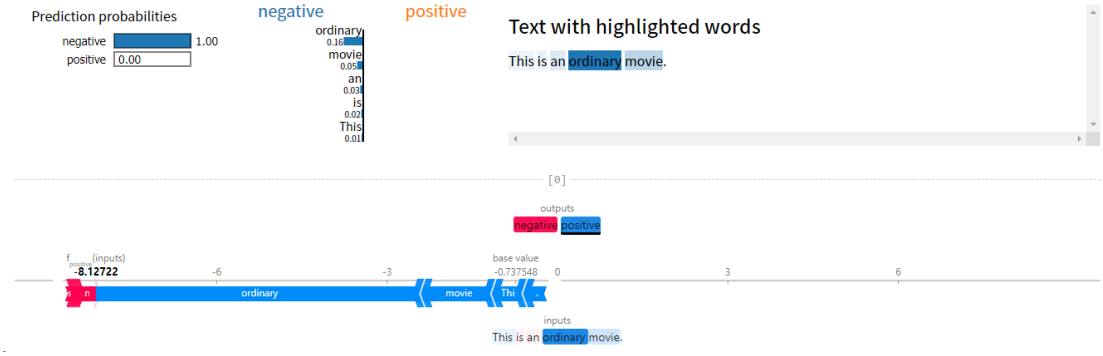


Figure 27: The explanation of first example given by LIME and SHAP.

And the second example I want to show is "This is a movie about woman". This example doesn't have obvious sentiment in it, too. But if we look at Figure 28, the model says that it's negative. To my surprise, if we change "woman" with "man" or "doctor" as in Figure 29 and Figure 30. The model think that this sentence becomes positive. I also don't know why this attack will work. In theory, model should treat these sentence as same because it doesn't have significant sentiment in it. But in reality, it's not what model works. To solve this problem, I think the datasets should add more sentences like these.

4.2 Misspelling

As we can see in Figure 31, if we misspell "bad" as "had", our model thinks that it has 0.77 probability to be positive. Therefore, I think that it's a successful attack.

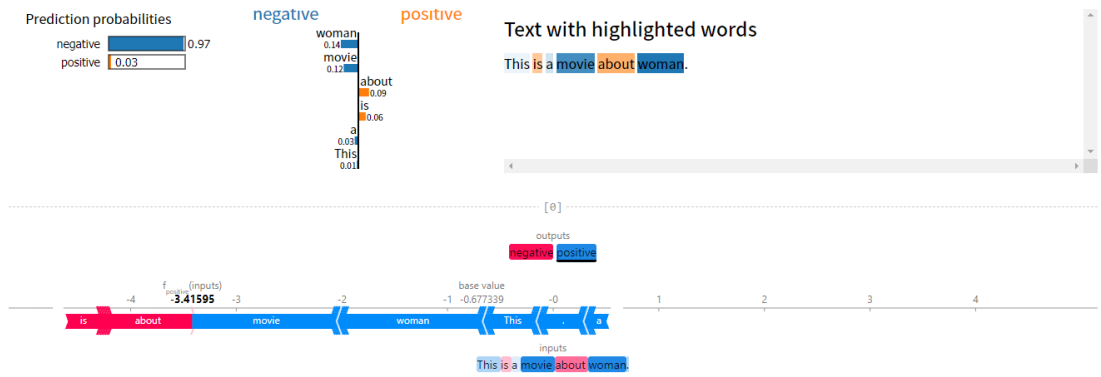


Figure 28: The explanation of second example given by LIME and SHAP.

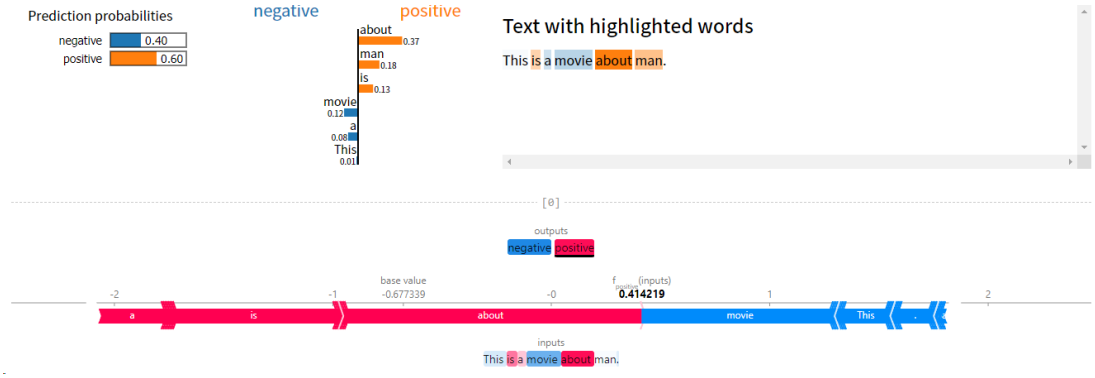


Figure 29: Changing “woman” to “man”.

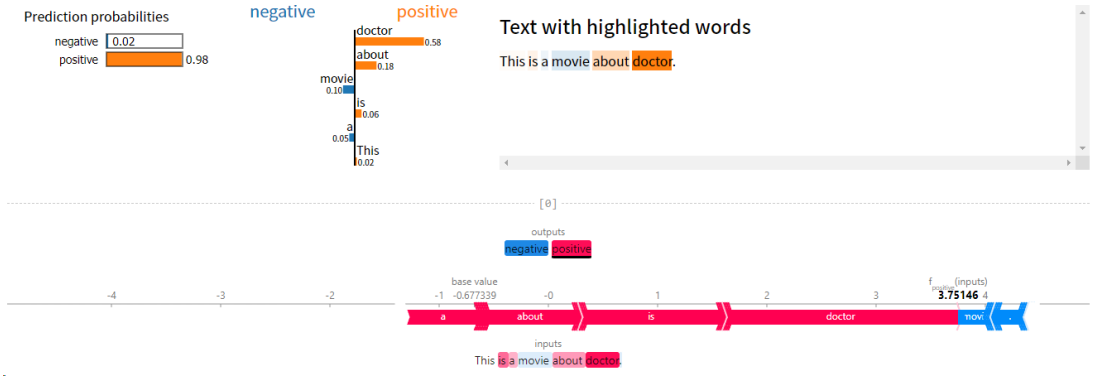


Figure 30: Changing “woman” to “doctor”.

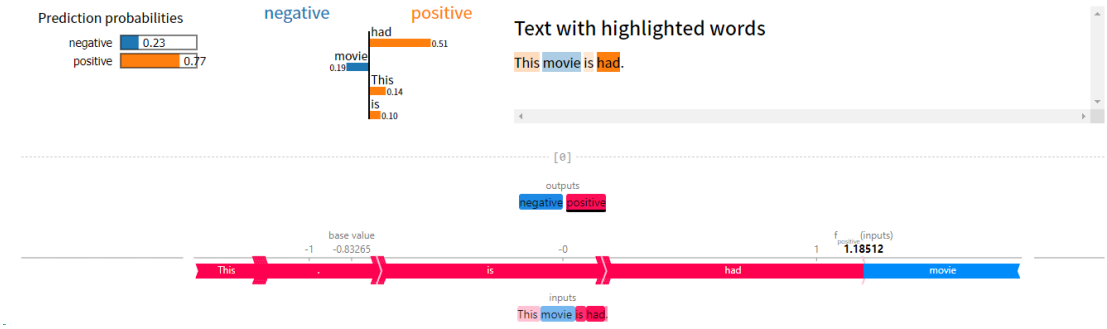


Figure 31: Using the sentence "This movie is had." to attack BERT.

Another using misspelling to attack BERT is shown as Figure 33. I misspell "good" as "goood". Although I only add a single "o" in good. The output of BERT change from 100% positive to 100% negative. Therefore, it's also a successful attack.

Figure 32: The original sentence.

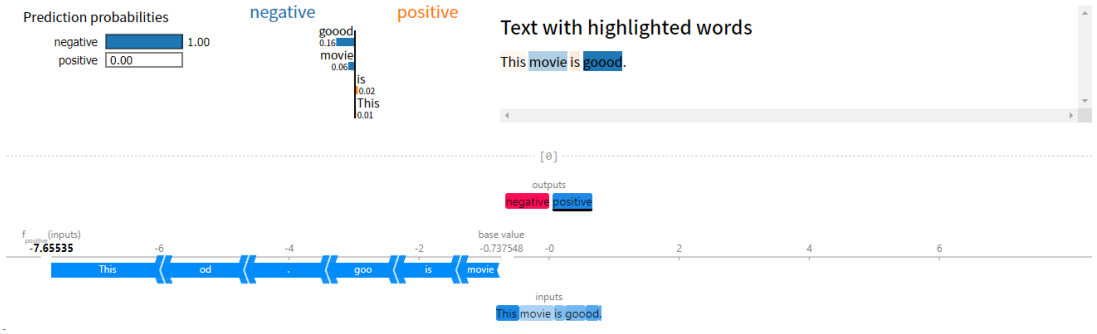


Figure 33: Misspell "good" as "goood".

4.3 Other method

The strategy I use in Figure 34 to attack BERT is to put as many negatives word as I can in this sentence, but the final sentiment of this review is positive. As we can see, this strategy works. The model thinks that 91% change that this sentence is negative.

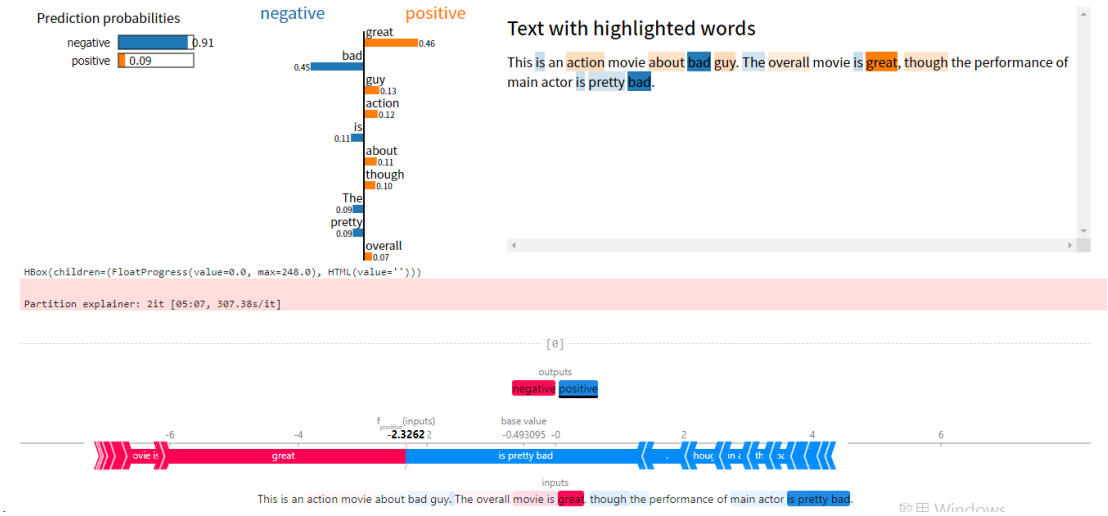


Figure 34: Using the sentence "This is an action movie about bad guy. The overall movie is great, though the performance of main actor is pretty bad." to attack BERT.

4.4 Using `textattack` package

"TextAttack is a Python framework for adversarial attacks, data augmentation, and model training in NLP."[5]

We can use this package to attack BERT model. For IMDB dataset, TextAttack has provided us the sample code. This code will first use `transformers` package to download pretrained model and IMDB datasets, then it randomly picks 20 samples to attack. And the attack strategy used here is `TextFoolerJin2019`, this is the method from "Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment"[6]. This strategy is to swap the word with 50 closest embedding nearest neighbors. This method successfully attack 19 sample in 20 sample. Below are two samples that attack successfully.

Code 1: Code for attacking BERT

[5]From its Github `https://github.com/QData/TextAttack`
[6]`https://arxiv.org/abs/1907.11932`

```python
import transformers
import textattack
model = transformers.AutoModelForSequenceClassification.from_pretrained("textattack/bert-base-uncased-imdb")
tokenizer = transformers.AutoTokenizer.from_pretrained("textattack/bert-base-uncased-imdb")
model_wrapper = textattack.models.wrappers.HuggingFaceModelWrapper(model, tokenizer)

dataset = textattack.datasets.HuggingFaceDataset("imdb", split="test")
attack = textattack.attack_recipes.TextFoolerJin2019.build(model_wrapper)
# Attack 20 samples with CSV logging and checkpoint saved every 5 interval
attack_args = textattack.AttackArgs(num_examples=20, log_to_csv="log.csv", checkpoint_interval=5, checkpoint_dir="checkpoints", disable_stdout=True)
attacker = textattack.Attacker(attack, dataset, attack_args)
attacker.attack_dataset()
```

Original Score	Perturbed Score	Original Score	Perturbed Score
I went and [[saw]] this movie last night after being coaxed to by a few friends of mine. I'll admit that I was reluctant to see it because from what I knew of Ashton Kutcher he was only able to do [[comedy]]. I was wrong. Kutcher played the character of Jake Fischer very well, and Kevin Costner played Ben Randall with such professionalism. The sign of a [[good]] movie is that it can toy with our [[emotions]]. [[This]] one [[did]] exactly that. The entire [[theater]] (which was sold out) was overcome by laughter during the first half of the movie, and were moved to tears during the second half. While exiting the theater I not only saw many women in tears, but many full grown men as well, trying desperately not to let anyone see them crying. This movie was great and I suggest that you go see it before you judge.	I went and [[endured]] this movie last night after being coaxed to by a few friends of mine. I'll admit that I was reluctant to see it because from what I knew of Ashton Kutcher he was only able to do [[travesty]]. I was wrong. Kutcher played the character of Jake Fi scher very well, and Kevin Costner played Ben Randall with such professionalism. The sign of a [[adequate]] movie is that it can toy with our [[sentiments]]. [[These]] one [[suis]] exactly that. The entire [[filmmaking]] (which was sold out) was overcome by laughter during the first half of the movie, and were moved to tears during the second half. While exiting the theater I not only saw many women in tears, but many full grown men as well, trying desperately not to let anyone see them crying. This movie was great, and I suggest that you go see it before you judge.	0.000188529	0.979928315
I saw this film on September 1st, 2005 in Indianapolis. I am one of the judges for the Heartland Film Festival that screens films for their Truly Moving Picture Award. A Truly Moving Picture "...explores the human journey by artistically expressing hope and respect for the positive values of life." Heartland gave that award to this film. This is a story of golf in the early part of the 20th century. At that time, it was the game of upper class and rich "gentlemen", and working people could only participate by being caddies at country clubs. With this backdrop, this based-on-a-true-story unfolds with a young, working class boy who takes on the golf establishment and the greatest golfer in the world, Harry Vardon. And the story is inspirational. Against all odds, Francis Ouimet (played by Shia LaBeouf of "Holes") gets to compete against the greatest golfers of the U.S. and Great Britain at the 1913 U.S. Open. Francis is ill-prepared, and has a child for a caddy. (The caddy is hilarious and motivational and steals every scene he appears in.) But despite these handicaps, Francis displays courage, spirit, heroism, and humility at this world class event. And, we learn a lot about the early years of golf; for example, the use of small wooden clubs, the layout of the short holes, the manual scoreboard, the golfers swinging with pipes in their mouths, the terrible conditions of the greens and fairways, and the play not being canceled even in torrential rain. This film has [[stunning]] cinematography and art direction and editing. And with no big movie stars, the story is somehow more believable. This adds to the inventory of [[great]] sports movies in the vein of "Miracle" and "Remember the Titans." FYI - There is a Truly Moving Pictures web site where there is a listing of past winners going back 70 years.	I saw this film on September 1st, 2005 in Indianapolis. I am one of the judges for the Heartland Film Festival that screens films for their Truly Moving Picture Award. A Truly Moving Picture "...explores the human journey by artistically expressing hope and respect for the positive values of life." Heartland gave that award to this film. This is a story of golf in the early part of the 20th century. At that time, it was the game of upper class and rich "gentlemen", and working people could only participate by being caddies at country clubs. With this backdrop, this based-on-a-true-story unfolds with a young, working class boy who takes on the golf establishment and the greatest golfer in the world, Harry Vardon. And the story is inspirational. Against all odds, Francis Ouimet (played by Shia LaBeouf of "Holes") gets to compete against the greatest golfers of the U.S. and Great Britain at the 1913 U.S. Open. Francis is ill-prepared, and has a child for a caddy. (The caddy is hilarious and motivational and steals every scene he appears in.) But despite these handicaps, Francis displays courage, spirit, heroism, and humility at this world class event. And, we learn a lot about the early years of golf; for example, the use of small wooden clubs, the layout of the short holes, the manual scoreboard, the golfers swinging with pipes in their mouths, the terrible conditions of the greens and fairways, and the play not being canceled even in torrential rain. This film has [[whopping]] cinematography and art direction and editing. And with no big movie stars, the story is somehow more believable. This adds to the inventory of [[sumptuous]] sports movies in the vein of "Miracle" and "Remember the Titans." FYI - There is a Truly Moving Pictures web site where there is a listing of past winners going back 70 years.	0.00015521	0.989287198

Figure 35: Two example that attack successfully. (words that has [[]] are the word being swapped)

5 Encountered Problems

5.1 Movie review contain both ' and "

When using long movie review to test model, the review may contain both ' and ". Therefore, we can't use normal "" or " to declare a string, instead, we will need to use triple quotes to create a string. This can also avoid some problem if the review has \n or other escape characters.

5.2 Can't find CUDA library

I run jupyter notebook on a workstation, and I don't know why notebook doesn't get the installation path of CUDA. Therefore, I just add `!export LD_LIBRARY_PATH="/usr/local/cuda/lib64/"` to set environment variable before importing `textattack` package. And this solve the problem I encounter.

5.3 Can't use `textattack` package to attack TA's model directly

If I use TA's model to create a model wrapper and try to attack this model. The error `RuntimeError: Could not infer dtype of tokenizers.Encoding` occurs. I don't know the meaning of this error. But I guess that it may relate to datasets. Because the datasets that is going to use to attack the model is loaded via `transformers` package. Therefore, it may be incompatible with the model TA gives.