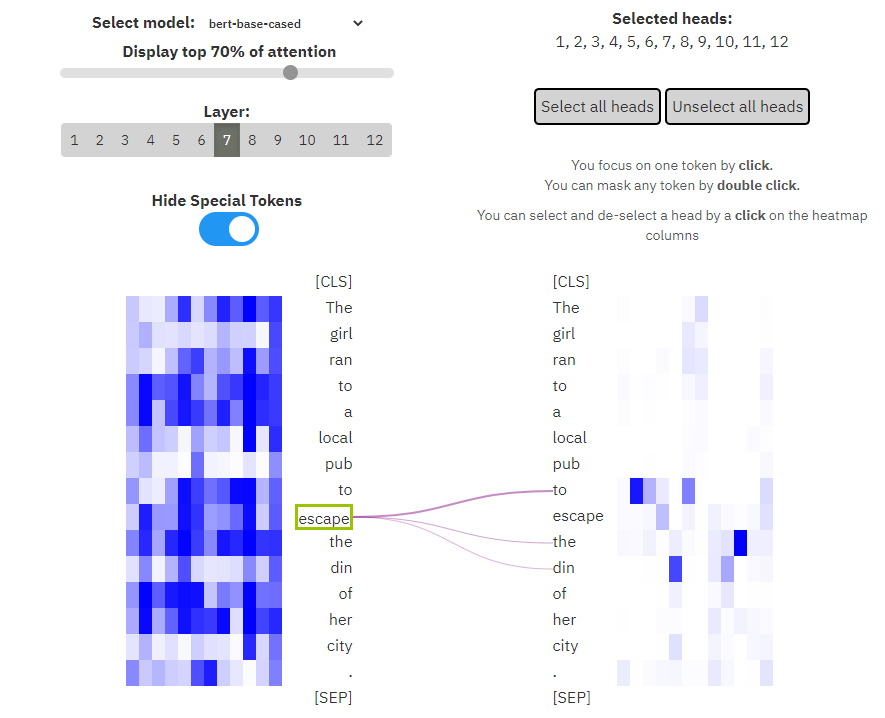
Attention Mechanism of BERT

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 使用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。

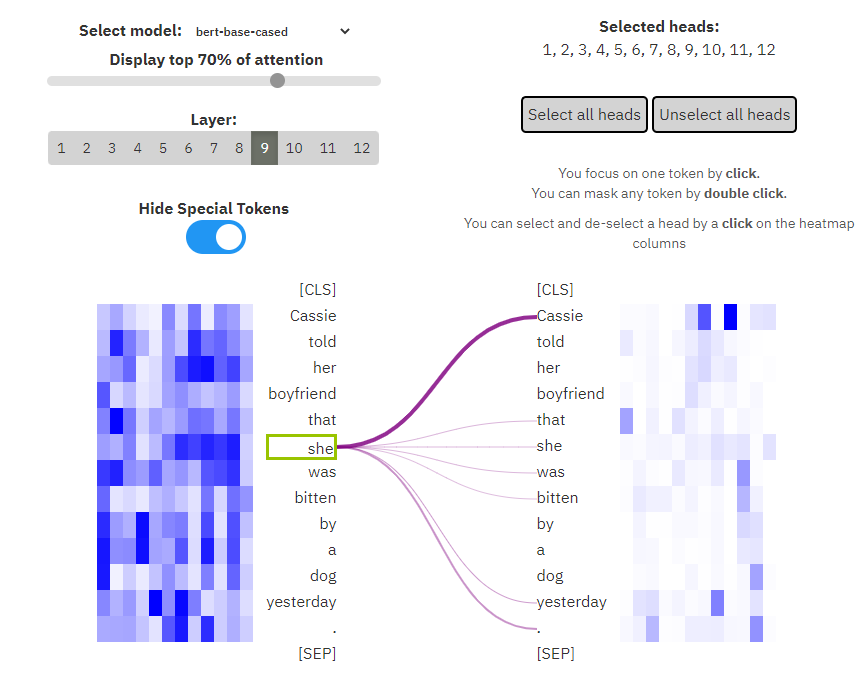


Language Modal：bert-base-cased

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。



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。

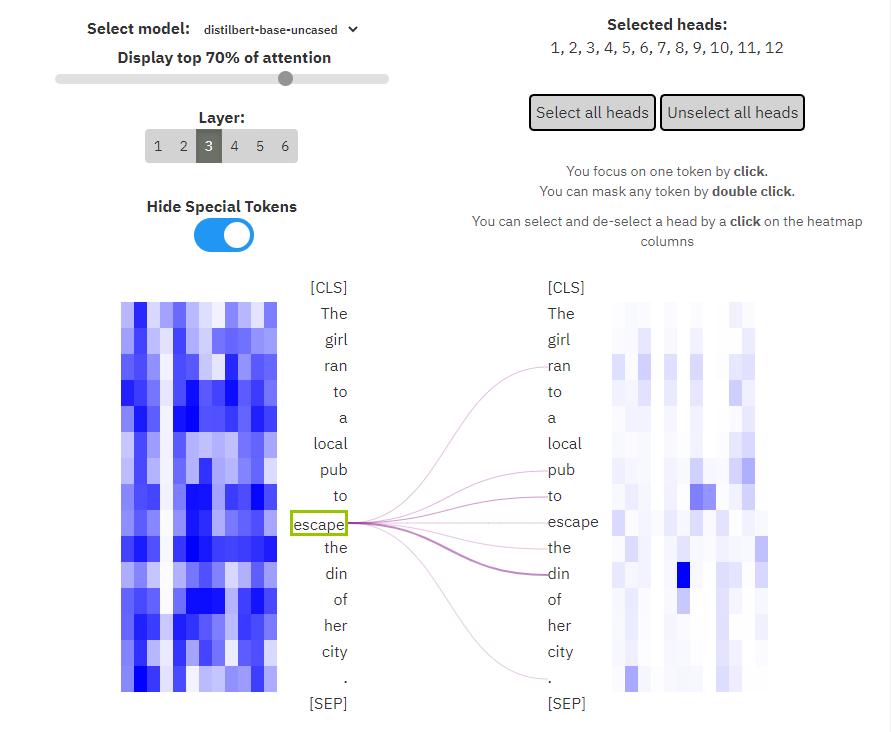
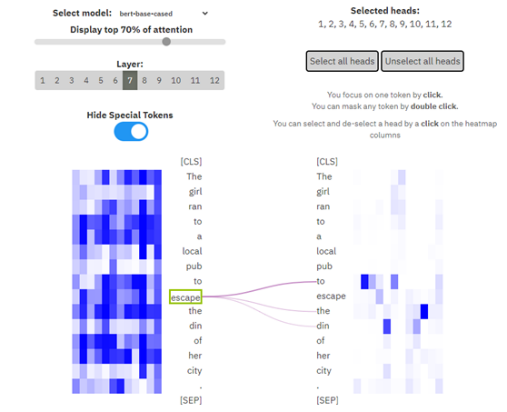
Attention 用於表示兩個詞語的關聯程度為何，以上兩個例子也說明了 attention mechanism 在 BERT 中是如何運作的，這種方法的確可以使的 BERT 的表現更好。

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。

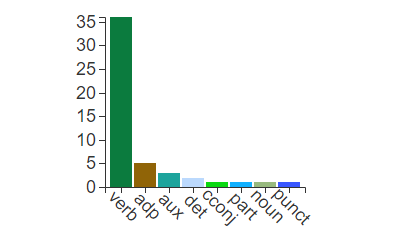
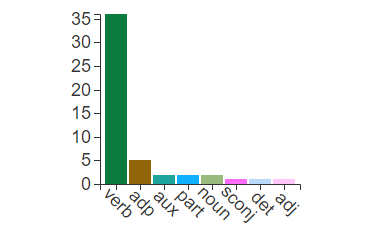
為了方便比較，我用相同的句子來測試，並遮住相同的詞語，句子如下：

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



bert-base-cased 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。

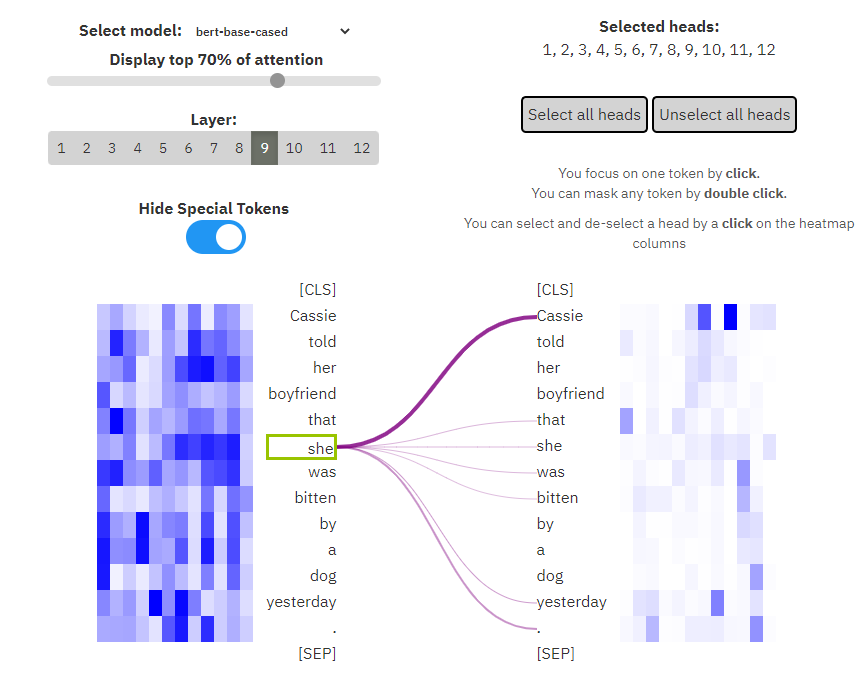
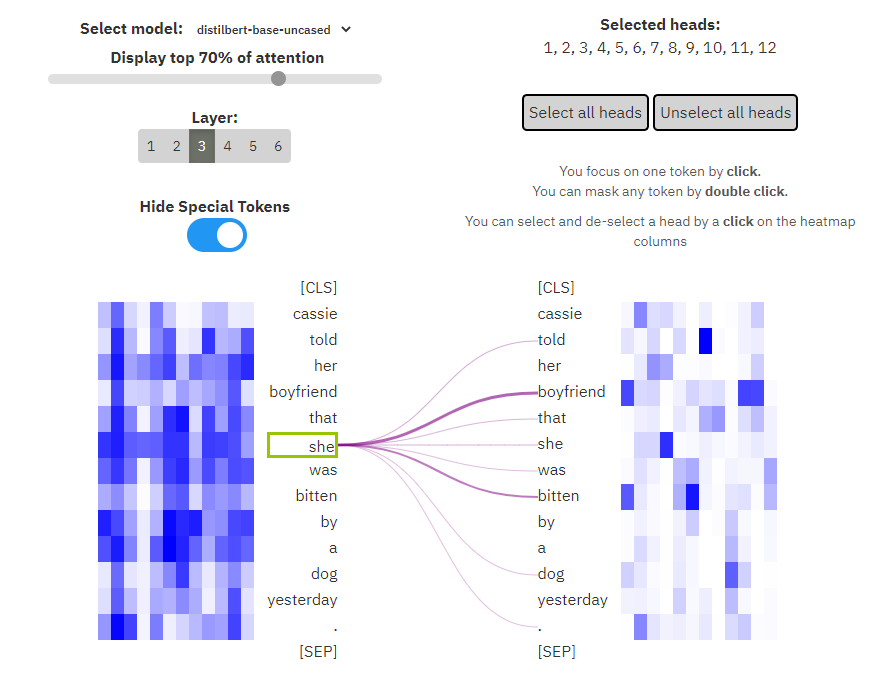


bert-base-cased distilbert-base-uncased

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。

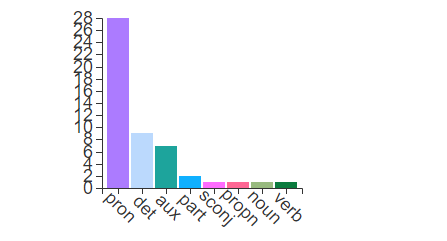
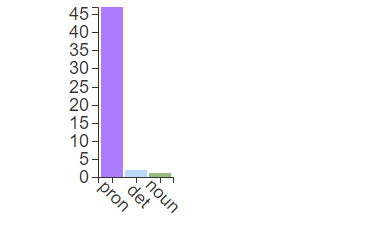
為了更精準的進行比較，我測試了另一個句子：

“Cassie told her boyfriend that she was bitten by a dog yesterday.”

bert-base-cased 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所預測的)。



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。

Compare of two sentiment classification models

3. Compare the explanation of **LIME and SHAP**. (30%)

4. Describe how you implement other explanation techniques. And discuss with the explanation result. (Bonus)

5. Try 3 different input sentences for **attacks**. Also, describe your findings and how to prevent the attack if you retrain the model in the future. (20%)

6. Describe problems you meet and how you solve them. (Bonus)