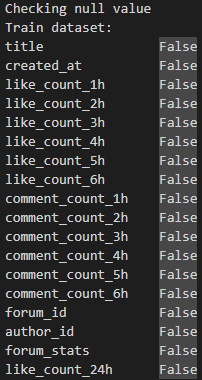
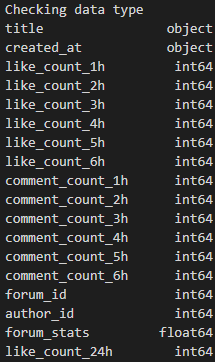
首先拿到資料集的第一件事，就是要先對資料有一定的了解及處理才可以開始做模型訓練，所以我先對資料集的每一個feature去檢查是否有遺失值的存在，得到的結果如下圖。



確認資料集皆沒有遺失值後，再來了解每一個feature的data type為何，以幫助我在選擇模型及資料預處理上有大致方向，結果如下。



對資料集有一定了解後，我選擇將“created\_at”、”forum\_id”、”author\_id”和”forum\_stats”這四項與預測結果相關度不高的features移除，並且因為title的型別為string，要可以與其他features併入一起做訓練的話，要將string轉換成數字的型別，如int或float。此時，我想到我學過的transformer的觀念，我決定將title去做word embedding將其轉換成數字vector再與其他features串接做成訓練集及驗證集。

根據網上查找到的資料，我選擇利用pretrained好的Bert model去做tokenization，並利用model的output取得embedding的結果，每一個title固定皆為768維度。