**BERT ATTENTION MAPS AND OBSERVATIONS**

Note:

In the following attention maps, the words in the left use the words in the right to make its embeddings ( outputs corresponding to its token ). The intensity of each of the 12 colors in the color spectrum of each word in the right represents the attention given by each of the 12 attention heads in BERT to that word.

Observations:

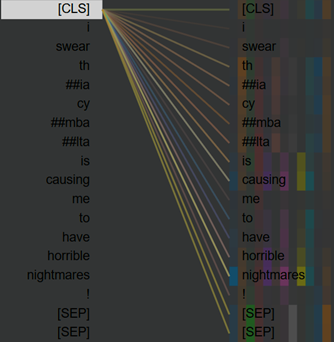
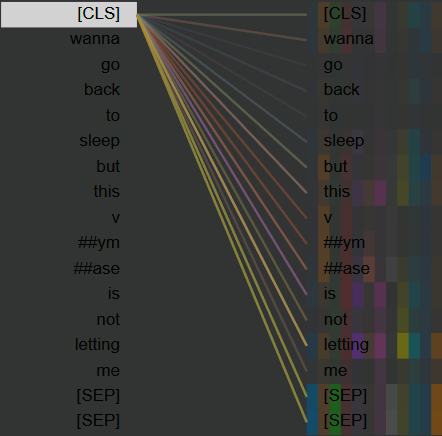
 

Figure : (1) “I swear thia cymbalta is causing me to have horrible nightmares!” ( Correctly predicted as 1)

(2) “wanna go back to sleep but this vymase is not letting me” (Correctly predicted as 1)

1. As can be seen in the above maps Bert pays attention to words representing user behavior/reaction in the vicinity to the name of the drug while also giving some attention to the drug name.

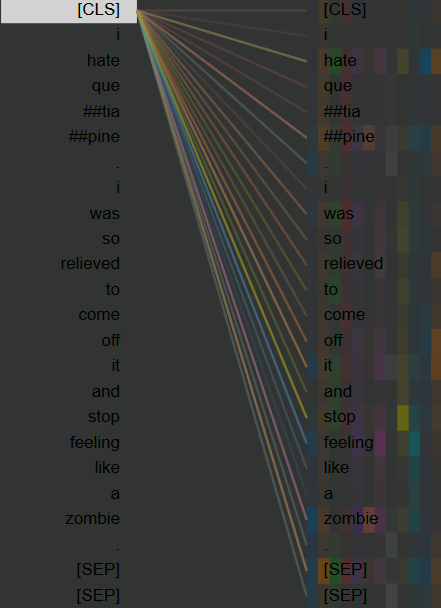
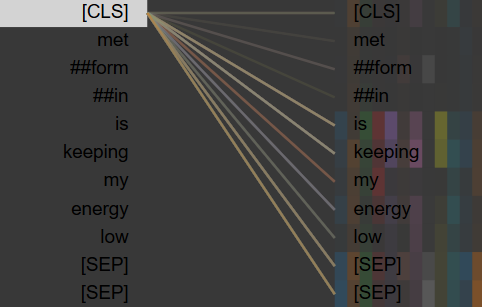


Figure : (3) “metformin is keeping my energy low” ( Incorrectly predicted as 0)

(4) “I hate quetiapine.I was so relieved to come off it and stop feeling lika a zombie. ” (Correctly predicted as 1)

2. Bert is more adapted to commonly occurring words like hate,help, hurt, feels, feeling, etc. and fails at seeing new words/phrases describing a reaction to self.

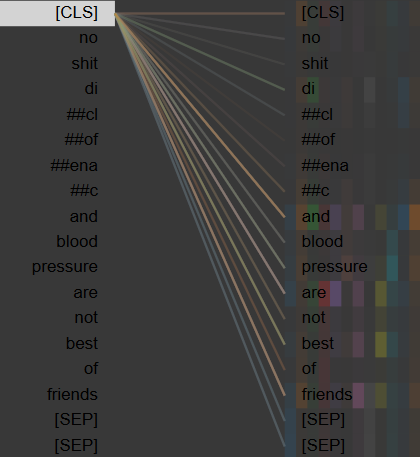
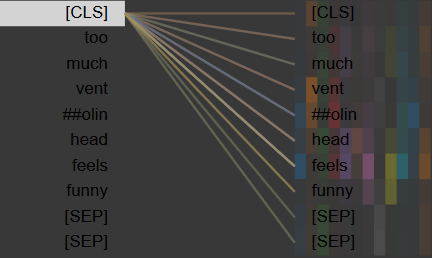
 

Figure : (5) “cipro is real bad on your entire body! take it from me on disability since,09 thanks to just 12 pills.” ( Incorrectly predicted as 0)

(6) “no shit diclofenac and blood pressure are not best of friends ” (Incorrectly predicted as 0)

3. The model is not good at capturing sarcasm and indirect mention of ADR i.e. containing no mention of any kind of harmful reaction to self.



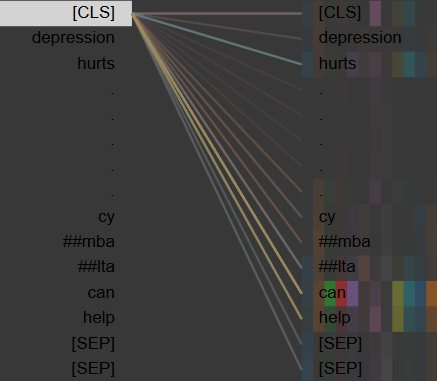
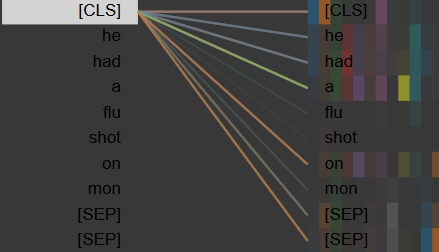
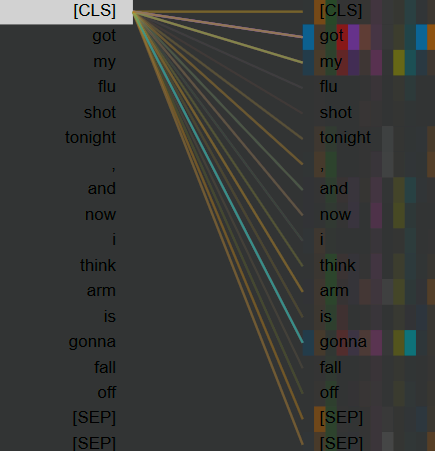


Figure : (7) “too much ventolin head feels funny” ( Correctly predicted as 1)

(8) “Depression hurts…..cymbalta can help ” (Correctly predicted as 0)

4.Bert can, however, identify the sentiment of the reaction i.e whether the reaction was good or bad(adverse).



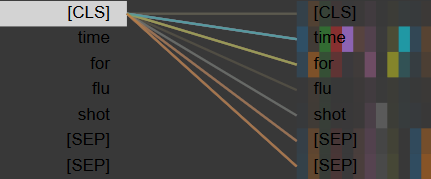
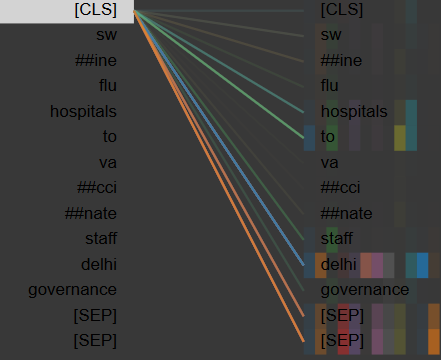


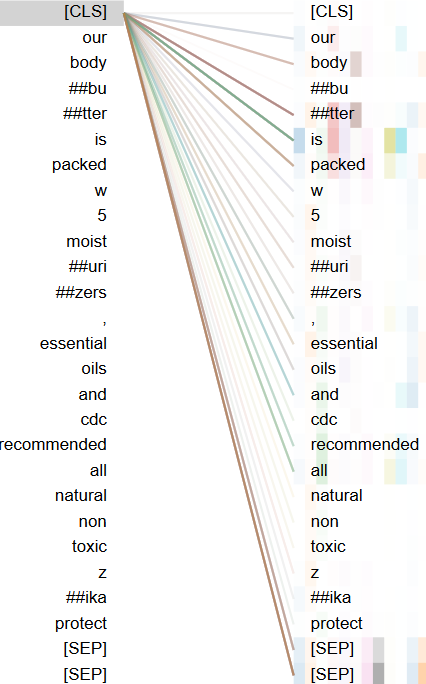
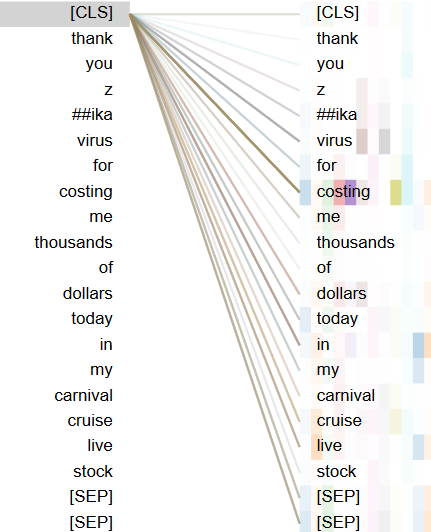
Figure (9): “Got my flu shot tonight, and now I think my arm is gonna fall off”(Correctly predicted as 1)

(10): “He had a flu shot on Mon”(Correctly predicted as 1)

(11) : “Time for flu shot”(Incorrectly predicted as 0)

(12): “Swine flu hospitals to vaccinate staff Delhi Governance”(Incorrectly predicted as 0)

Zika Specific Tweets-



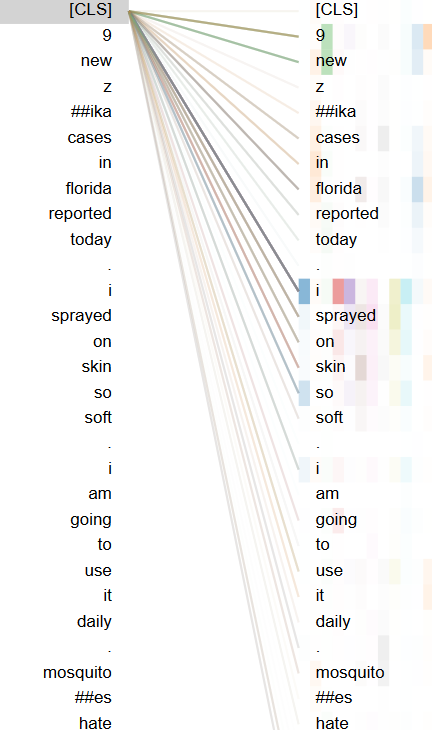
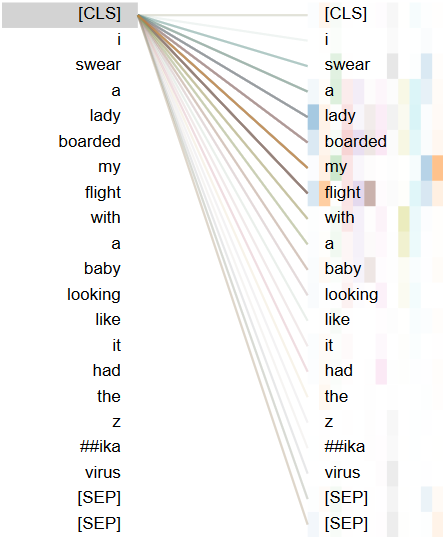


Figure (13): “Thank You Zika Virus for costing me thousands of dollars today in my Carnival Cruise live stock”(Incorrectly predicted as 1)

(14): “Our Bodybutter is packed w 5 moisturizers, essential oils and CDC recommended all natural non toxic Zika protect”(Incorrectly predicted as 0)

(15) : “9 new Zika cases in Florida reported today. I sprayed on Skin So Soft. I am going to use it daily. Mosquitoes hate...”(Correctly predicted as 1)

(16): “I swear a lady boarded my flight with a baby looking like it had the Zika virus”(Correctly predicted as 0)

Observations -

1.The model pays more attention to a combination of following POS tags to make its predictions:

1.Pronoun (me,my,I,he,she, etc.)

2.Verb ( need,needs,had,have,got,gonna,etc.)

3.Prepositions ( to,on,after,etc.)

2. Bert relies heavily on the above POS tags, which are responsible for signifying personal health, to make its predictions. Absence of words belonging to these categories almost always makes the model predict as **not** containing personal health mention.

3. Giving relatively less attention to words like ‘flu’, ‘shot’ and ‘zika’ shows both that the model is not overfitting those words and explains as to why Bert is performing so good in subtasks like C3 where there is no particular disease mention.

Note:I have included only the attention corresponding to the CLS token of the last layer in above maps as we are taking the output from the output from that only . The observations above are a combined result of analysis I did across many layers .

Link to this doc- <https://docs.google.com/document/d/1MADrJGK0u9rGb8y5oV1Z3WlfdD6rLJyedKk2mbm7S2k/edit?usp=sharing>