So this week, I am learning the paper professor Keith sent about SBERT. I found it's a really great model to enable sentence embeddings.

My idea is adding up all sentence vectors to find a story vector and find a linear combination of comment vectors to get the “evaluations” of this story vector. Finally I will add these vectors together to get the final vector for a post.

In particular, for example, here I made a panda table to show how it works. Here we have the tiltle of a post, its corresponding texts and scores. And Its comments and corresponding scores. I am using a groupby method, so this table shows that this post in particular has four comments with score 5, 3, 2, and 1.

Then using the sbert model from sentence_transformers with this pretrained model, one of the decent models the package provides. For the text, since the encoding is running on a sentence basis, so I summed up all the sentence embedding vectors to get a story vector of shape (768,). On the other hand, I did the same thing for each comment, but I then take a linear combination of comment vectors. For example, calculate the vector for these couple of sentences and multiply its score, and then calculate the vector for these couple of sentences and multiply its score, etc, etc. then add them all up to get a comment vector. Finally use text vector multiply its score and add the comment vector to get the final vector. This vector will be the features classification.

So the final result is this. Since the computational work is quite large, I am taking a relatively large subsets of the data. I ended up using half of the data per each subreddit. And I am also accelerating with my graphics card since spacy supports cuda. I have RTX 2070 here in my gaming pc and it greatly increased the performance by 4-5 times. My macbook only got 10-20 items per second, but with a gpu, it can get up to 60-70. Finally I have around 1600 instances for r/nosleep and 1700 for r/confessions.

Last but not least, input them into keras Neural network model. This is my first time using tensorflow. I tried to use 1 faltten layer and 3 dense layers, but it doesn't come up very well. I guess I can improve this part next week.

This is what I did for now. Thank you for listening.