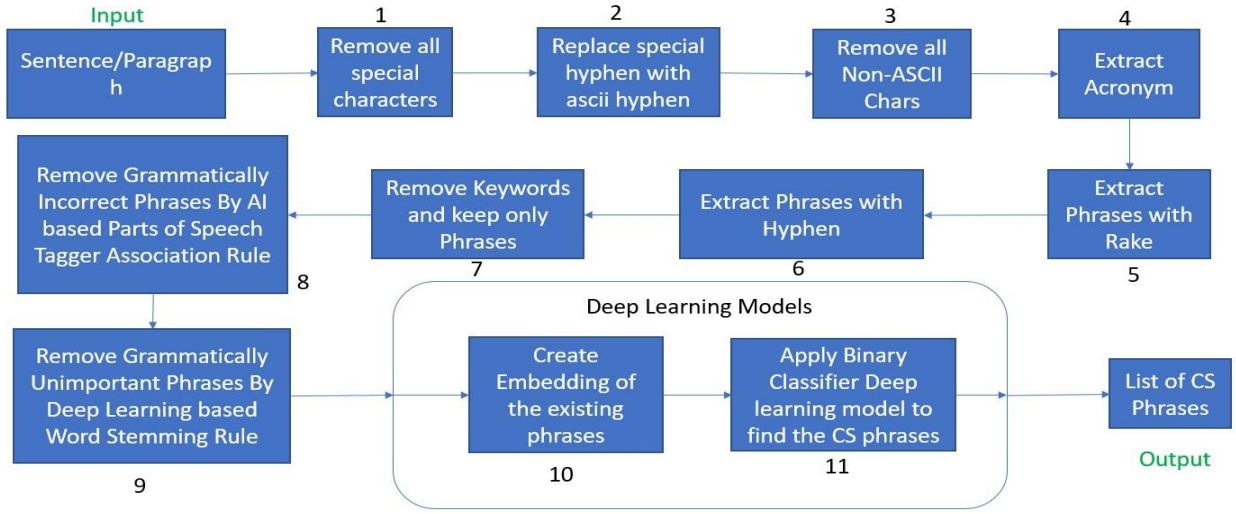
Automatic Computer Science Phrase Extraction: Model Visualization

S M Abrar Jahin

In our day-to-day life, we are using different search engines for searching different content. During the search, we need to express our problems with a keyword or phrase. Moreover, search engines are also providing us with suggestions as a form of type help which are also keywords and phrases. I am thinking of a better way of extracting phrases from any sentences (will also work on paragraphs or articles). When we start searching about that, then we found there are some existing algorithms and papers that are providing the same tasks. At that time, we think about doing a novel work, so we like to discretize our work from existing works which adds new dimensions to our work. As far as we have searched, we didn't find any work which is doing only phrase extraction on a specific domain.

For doing that, I am working on this algorithm-



In short, algorithm is looking like this-

1. Remove all special characters
2. Replace special hyphen with ascii hyphen
3. Replace the chars to ASCII char so that encoding special chars can be removed
4. Get Acronym List
5. Extract potential keywords and phrases with RAKE
6. Add hyphen related keywords from before word and after word, Then add hyphen related phrases from before keyword and after keyword from the found phrase list
7. Remove keywords and keep only phrases
8. Remove keywords (words and phrase) based on Parts of Speech Tagger (Determinants [a,an,the], Proper Nouns [person names], WH pronouns, predeterminer [all the kids], pronouns, verbs) -> Rule based classifier
9. Remove duplicate phrases by word stemming
10. Get embedding from pre-trained BERT model for all the phrase list
11. Remove CS phrases from the list with a pre-trained binary classifier

I have created a **dashboard** with different kind of interactive ways to visualize the deep learning models used in the algorithm.

Used Data Collection-

The data used in this project are collected from Wikipedia data extraction and labeled by hand by me. There are 2500+ data which are labeled, and they are looking like this-

902	virtual machine	cs	
903	virtual memory addresses	cs	
904	virtual reality	cs	
905	visual objects	cs	
906	visual programming languages	cs	
907	web caching	cs	
908	web design	cs	
909	web servers	cs	
910	weight-balanced tree	cs	
911	wireless routing protocol	cs	
912	world wide web	cs	
913	zip files	cs	
914	0 - sparse graphs	non-cs	
915	1 became important	non-cs	
916	1 - sparse graphs	non-cs	
917	1 - tight graphs	non-cs	
918	1 architectural view model	non-cs	
919	1 computer graphics system	non-cs	
920	1973 nbs solicited private industry	non-cs	
921	1975 distributed artificial intelligence emerged	non-cs	
922	1987 ai magazine article	non-cs	
923	1989 stochastic diffusion search	non-cs	
924	1996 maiden flight	non-cs	

And for model training data, I have made log in different step in different training for model and saved them in file after processing the logs for a better visualization.

Word Cloud-

Before going deep at the 6th step, I am getting keywords and phrases.

So, for visualizing them, I have created an animated word cloud which is looking like this-

The larger ones are the most frequent words and phrases, and smellers are less frequent keywords and phrases.

After getting the keywords and phrases, I like to do some processes which are discussed in algorithm from where at step 9, I am getting some list of phrases. But all of them are not computer science phrase. So, I need to create a system from where only CS phrases will remain. For doing that, I have used BERT. With BERT, I am getting an embedding of size 768. With those data, it is difficult to visualize the data. So to visualize that high dimensional data, I have used this 2 algorithms-

1. T-SNE
2. PCA

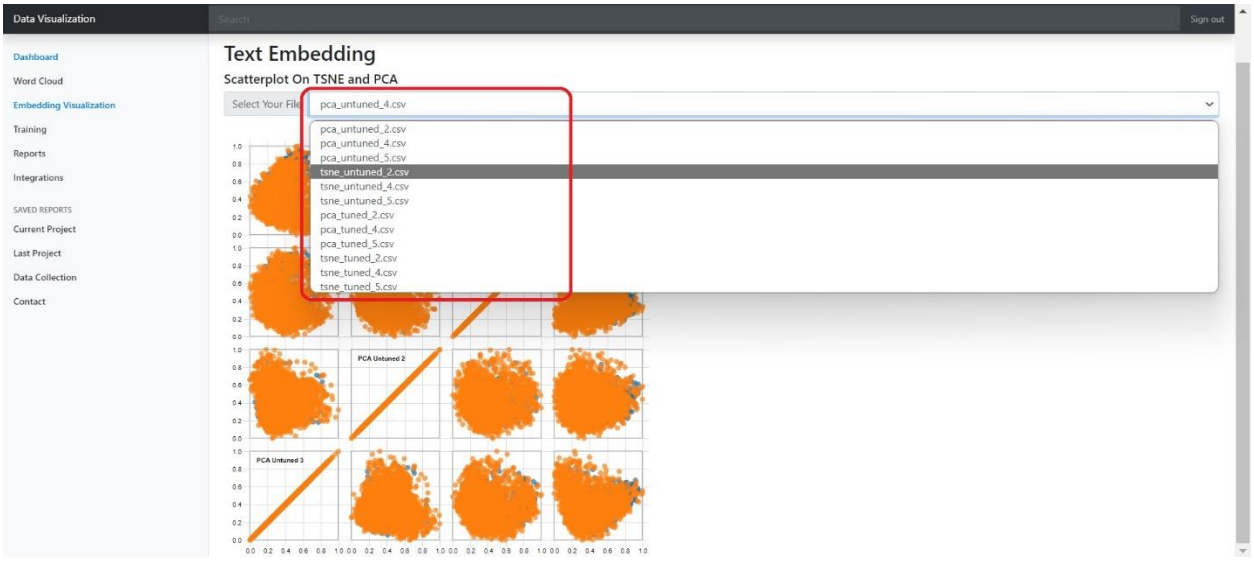
For easy visualization, I have created the data of 3 types of dimensions-

1. 2 dimensions
2. 4 dimensions
3. 5 dimensions

And I have created the visualization

1. Before BERT model training and
2. After BERT model training

My dashboard is looking like this-

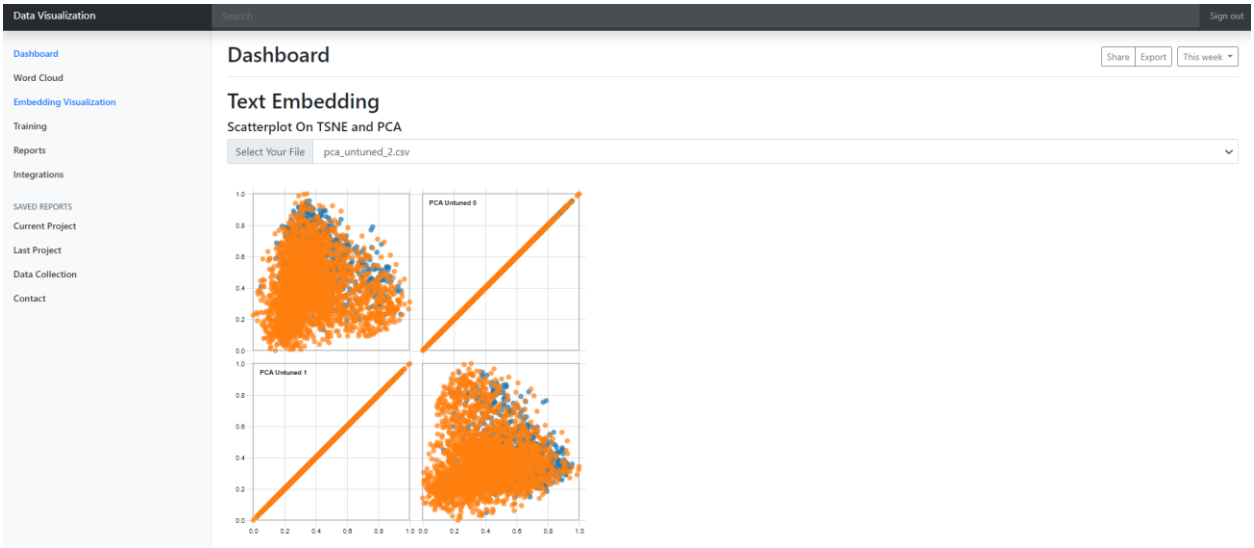


And all the figures are burstable like this-

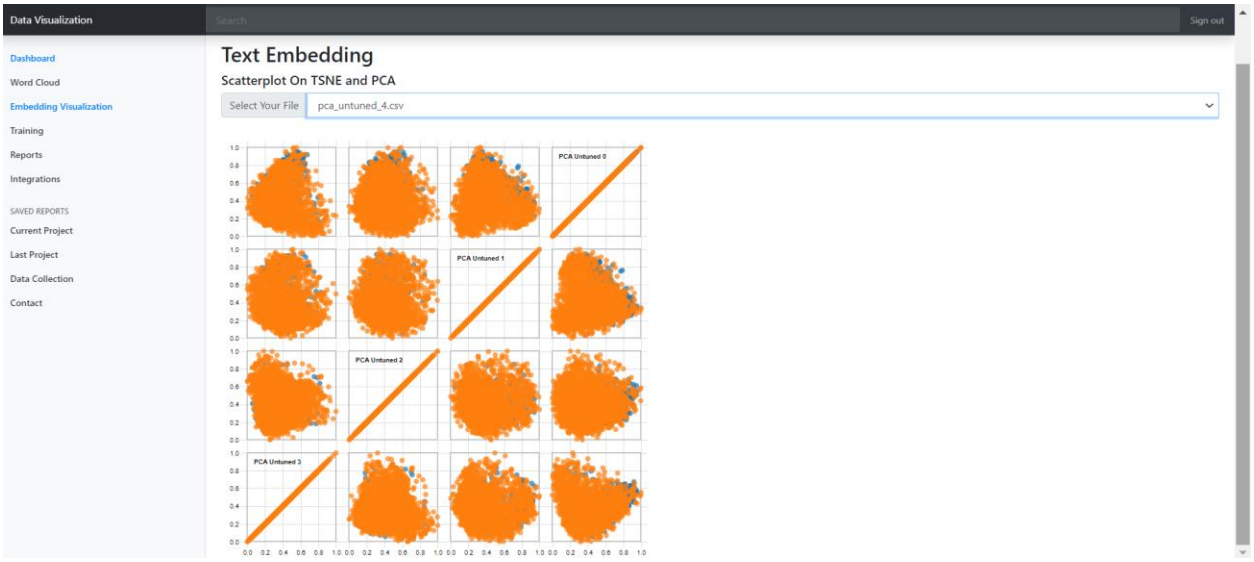
So, if we visualize PCA without BERT tuned model, I am getting data visualization like this-

PCA with Un-tuned BERT Model Embedding Data-

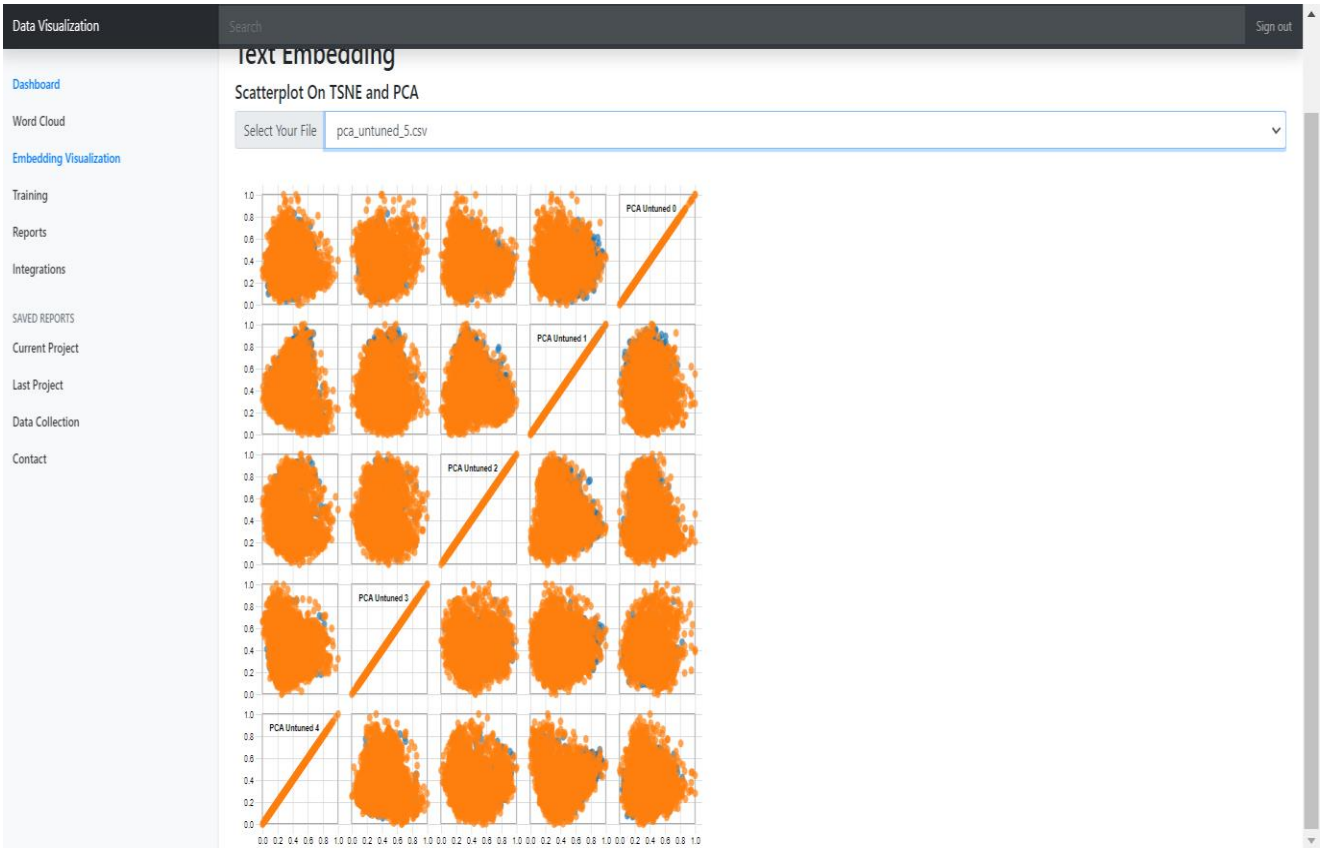
2 Dimension-



4 Dimension-

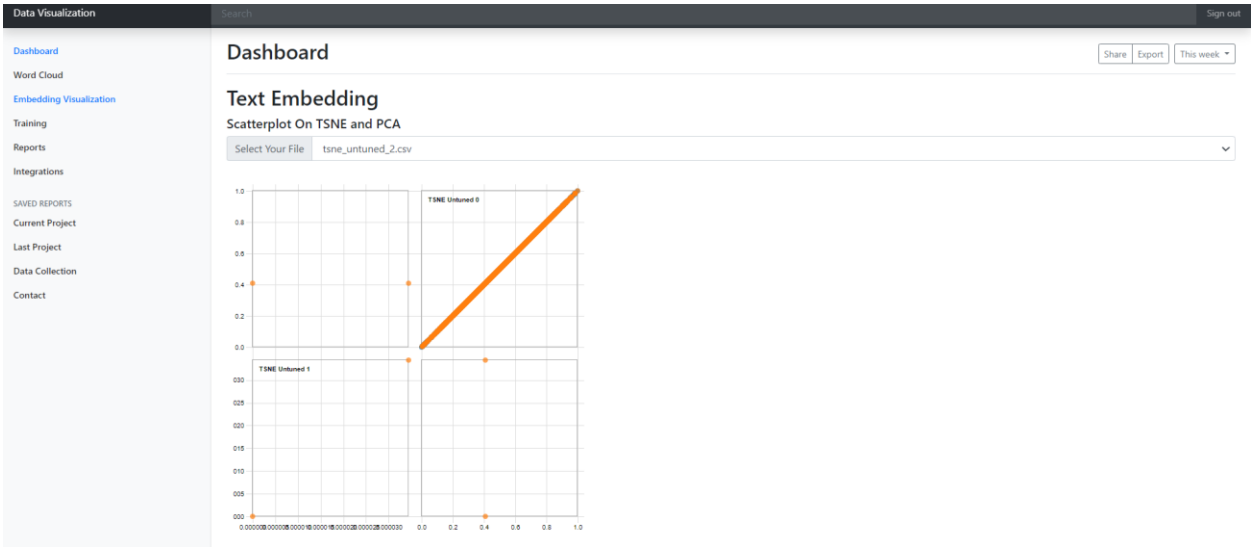


5 Dimension-

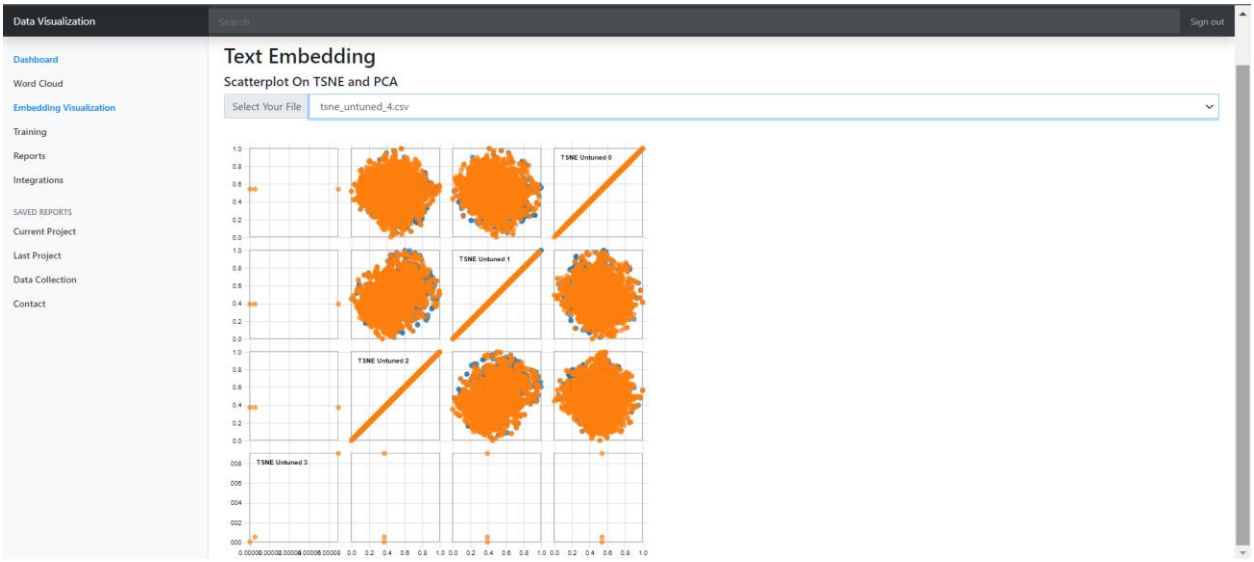

Data Visualization
Search
Sign out
Dashboard
Word Cloud
Embedding Visualization
Training
Reports
Integrations
SAVED REPORTS
Current Project
Last Project
Data Collection
Contact
Text Embedding
Scatterplot On TSNE and PCA
Select Your File
pca_untuned_5.csv
PCA Untuned 0
PCA Untuned 1
PCA Untuned 2
PCA Untuned 3
PCA Untuned 4

tSNE with Un-tuned BERT Model Embedding Data-

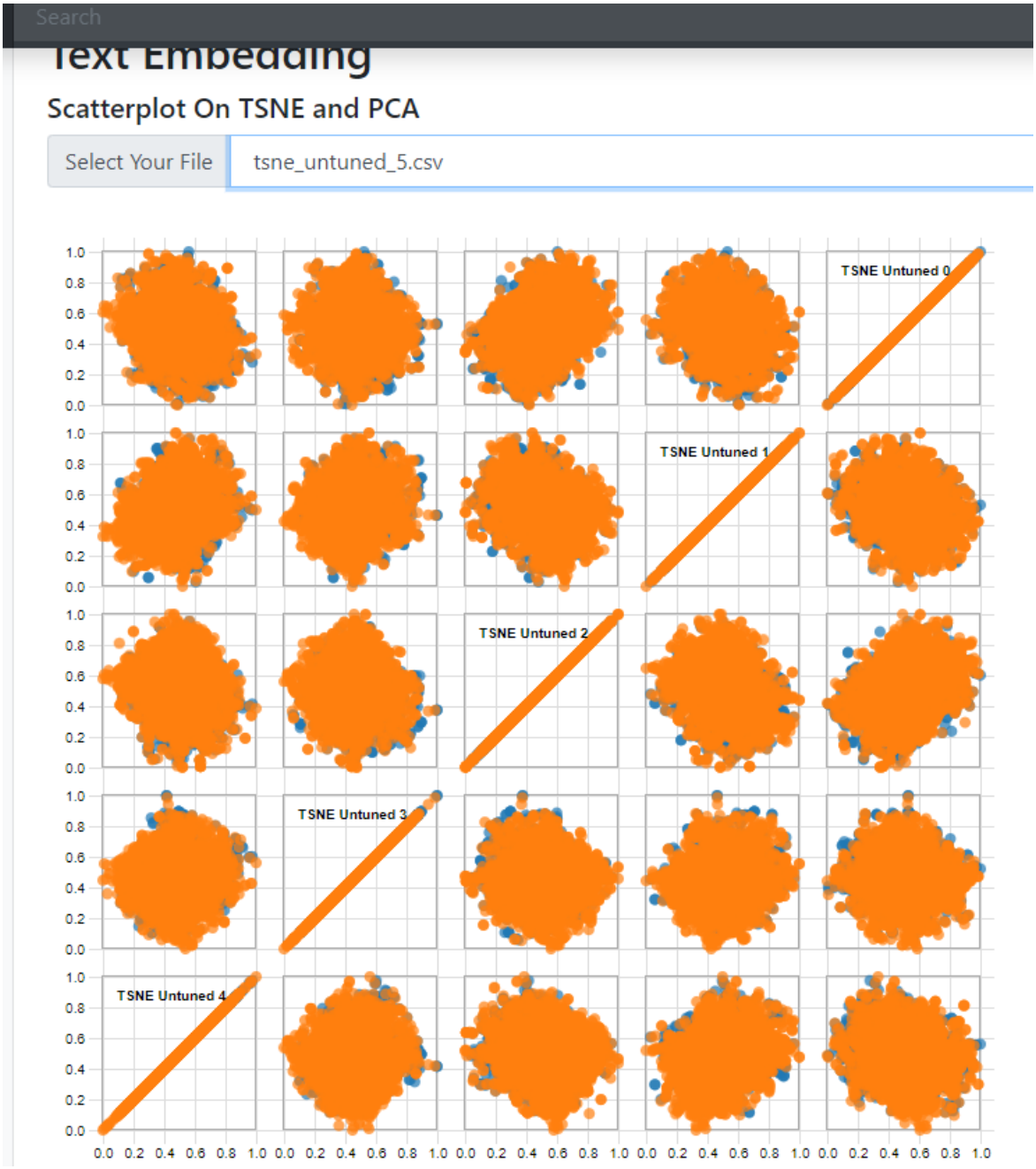
2 Dimension-



4 Dimention-



5 Dimention-

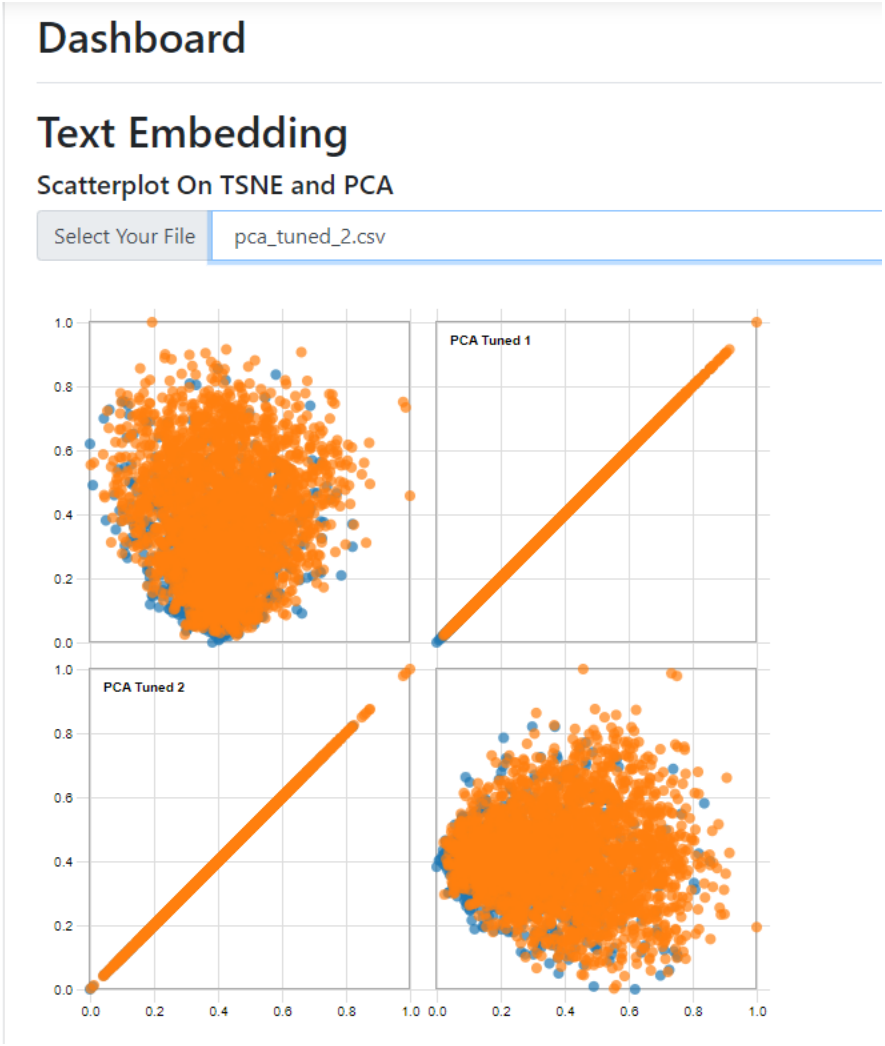


Se, we can see the data is not completely segregated.

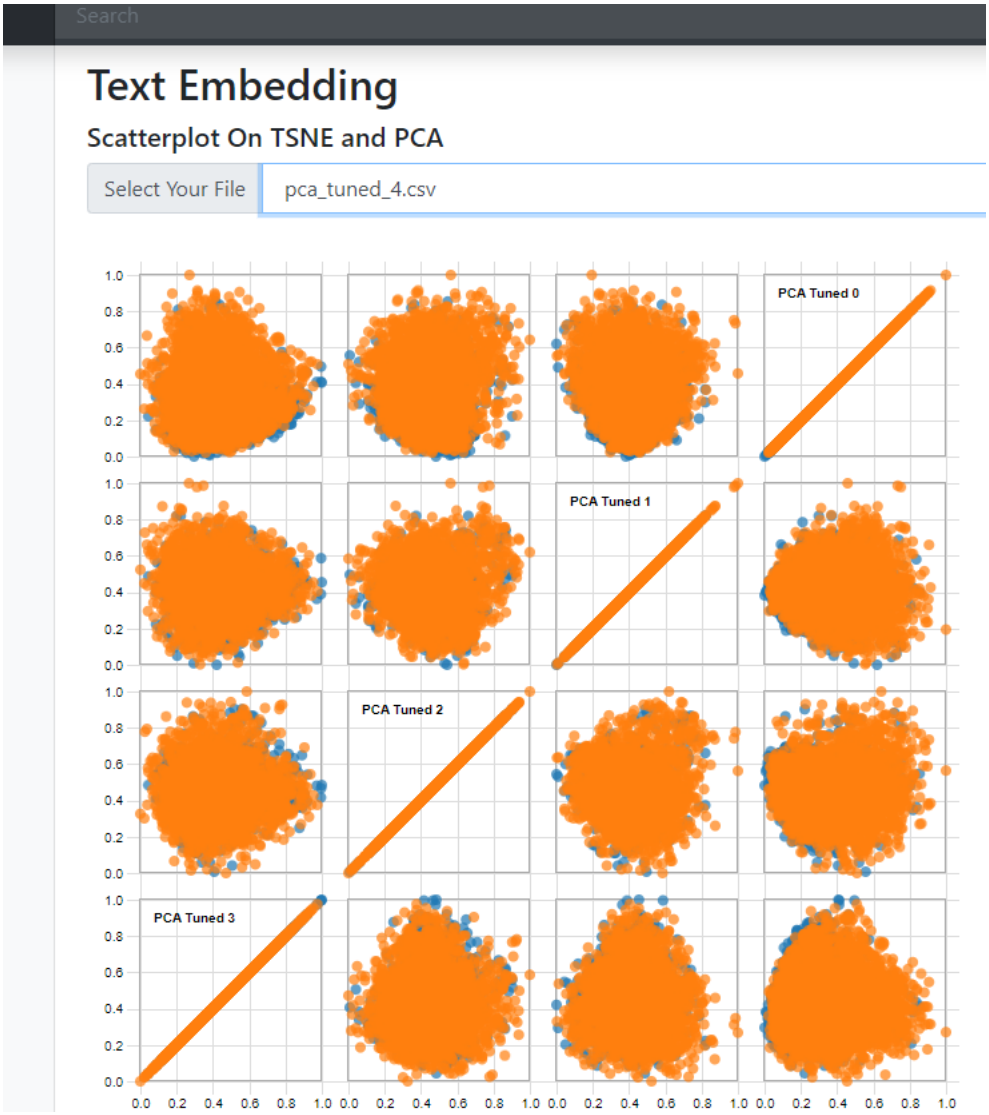
Now, let’s see what happens if we can tune the pre trained BERT.

PCA with tuned BERT Model Embedding Data-

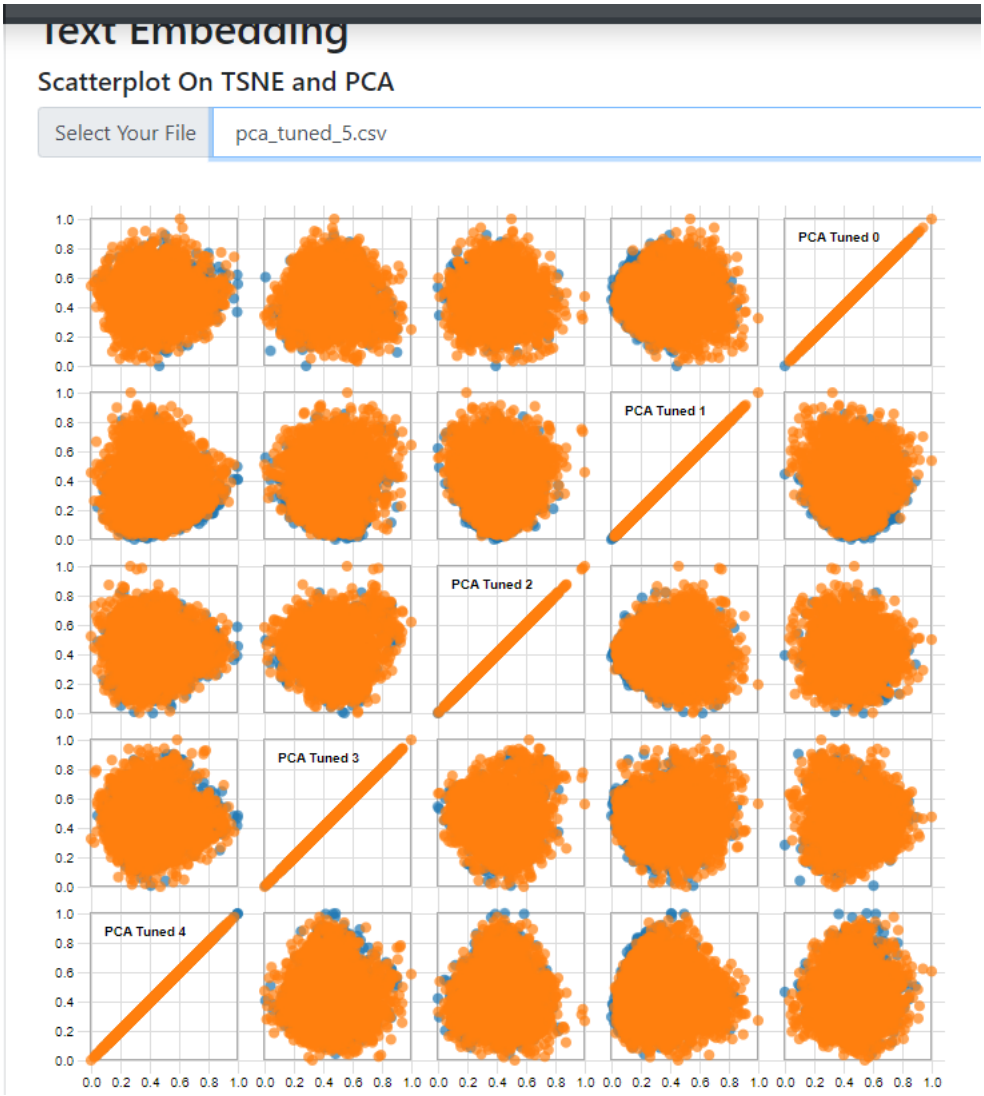
2 Dimension



4 Dimension

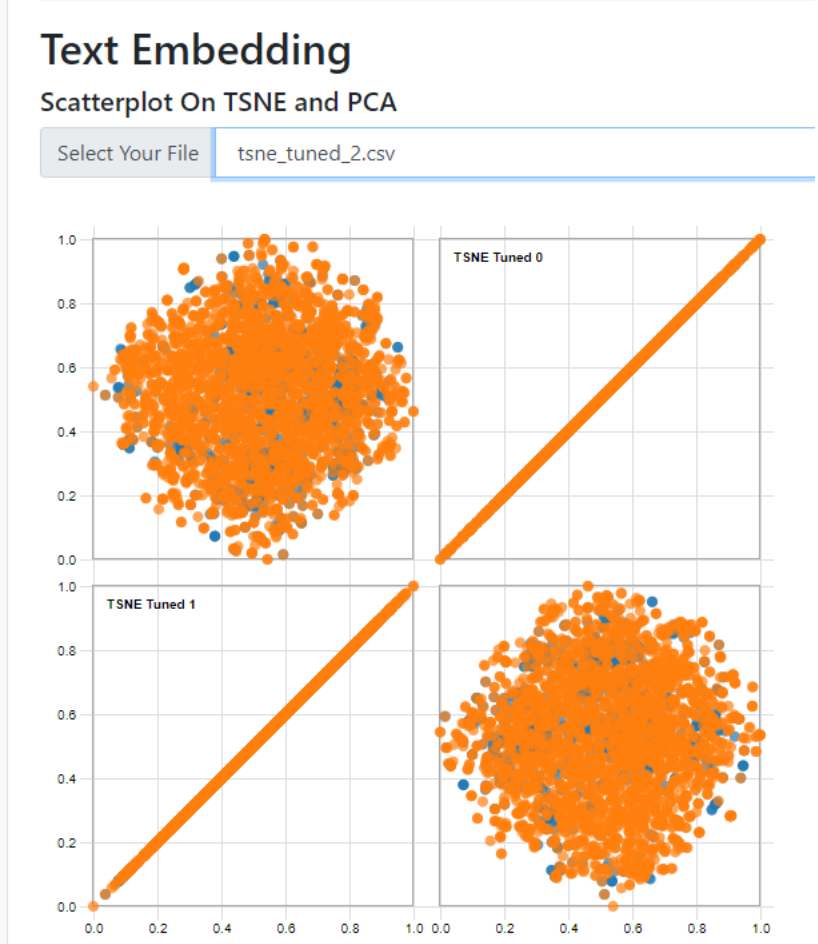

Search
Text Embedding
Scatterplot On TSNE and PCA
Select Your File
pca_tuned_4.csv
PCA Tuned 0
PCA Tuned 1
PCA Tuned 2
PCA Tuned 3
0.0
0.2
0.4
0.6
0.8
1.0

5 Dimention


Text Embedding
Scatterplot On TSNE and PCA
Select Your File
pca_tuned_5.csv
PCA Tuned 0
PCA Tuned 1
PCA Tuned 2
PCA Tuned 3
PCA Tuned 4

tSNE with tuned BERT Model Embedding Data-

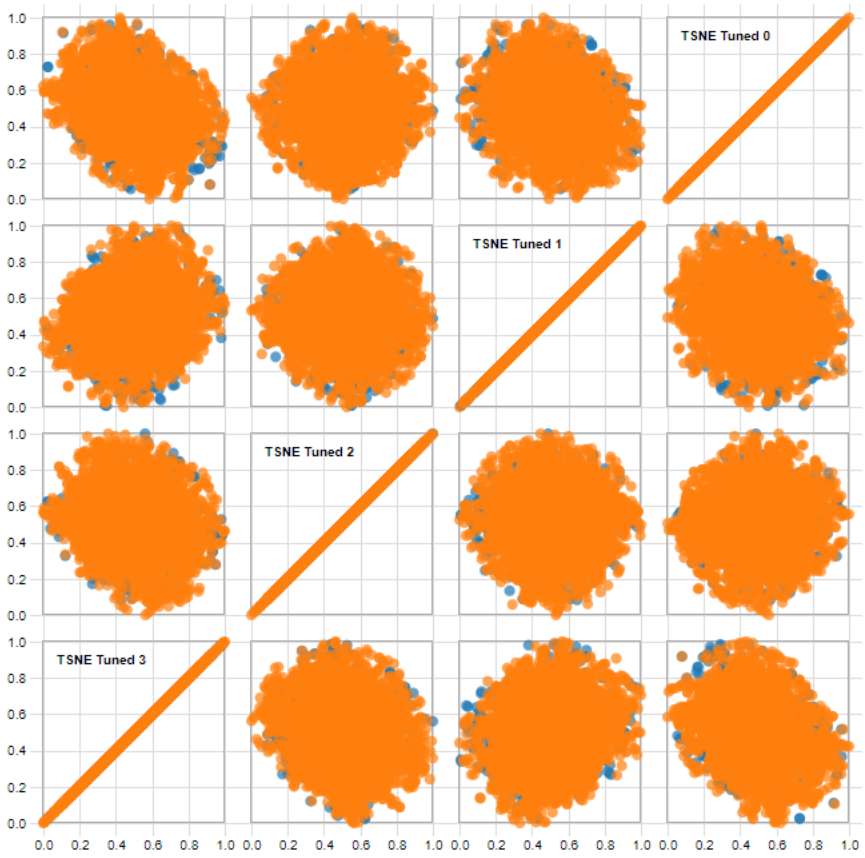
2 Dimension



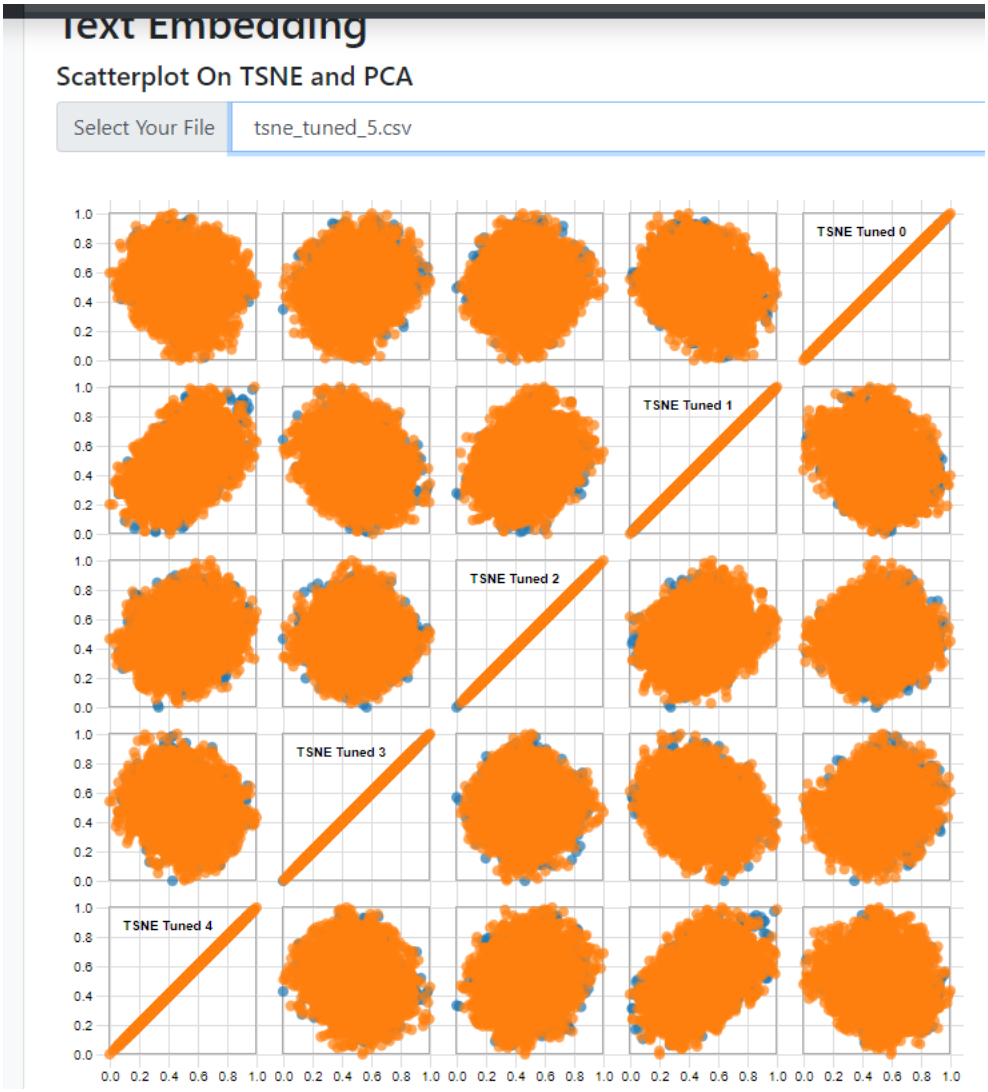
4 Dimension

Scatterplot On TSNE and PCA

Select Your File | tsne_tuned_4.csv



5 Dimention

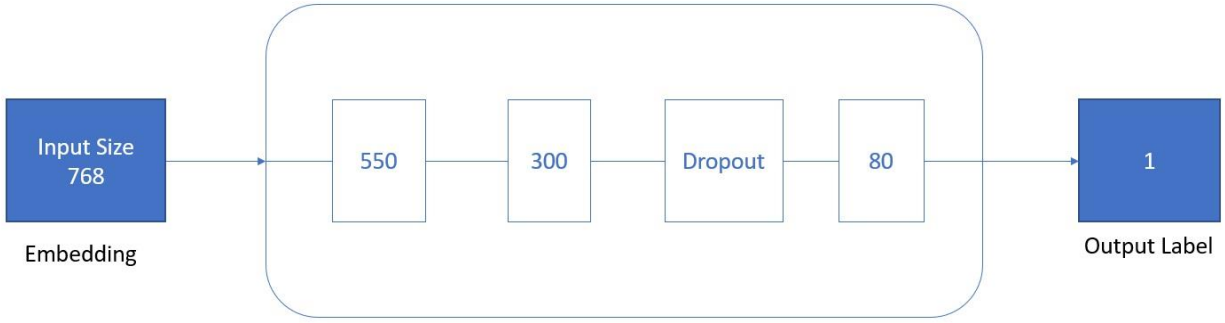


After analyzing those figures, we can see that, data are segregated, and it is seen that the orange-colored data are in center and blue colored data are distant from the center.

Analysis Result-

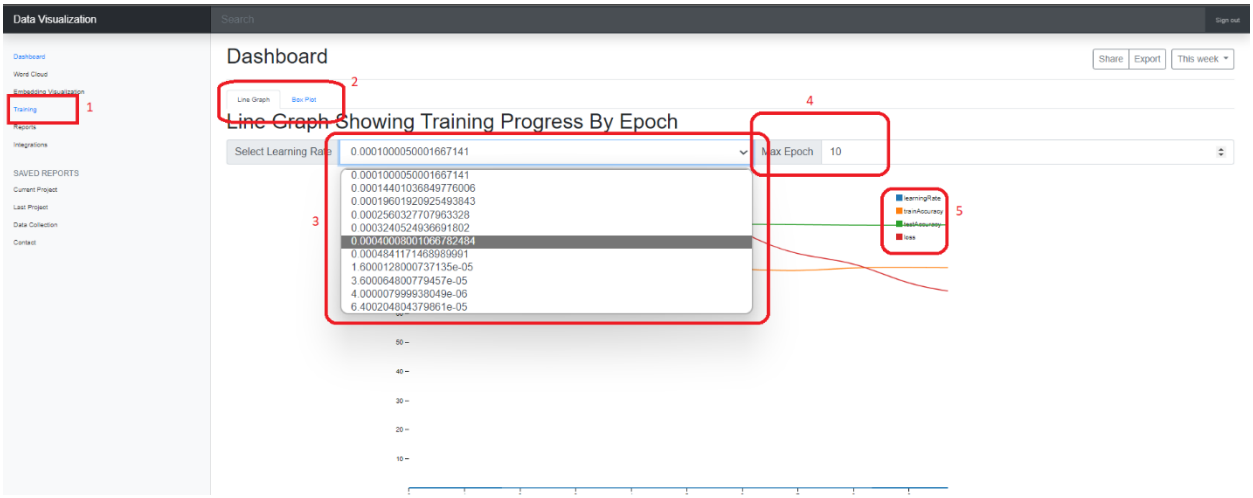
And **PCA is providing better visualizing than t-SNE** for my data which are collected from Wikipedia and labeled by in person by me and tuning the pretrained BERT is helping.

Now let's see what the binary classifier looks like mentioned in 11th step.



I have extracted the training data in different step of model by saving the data in log and extracting the log.

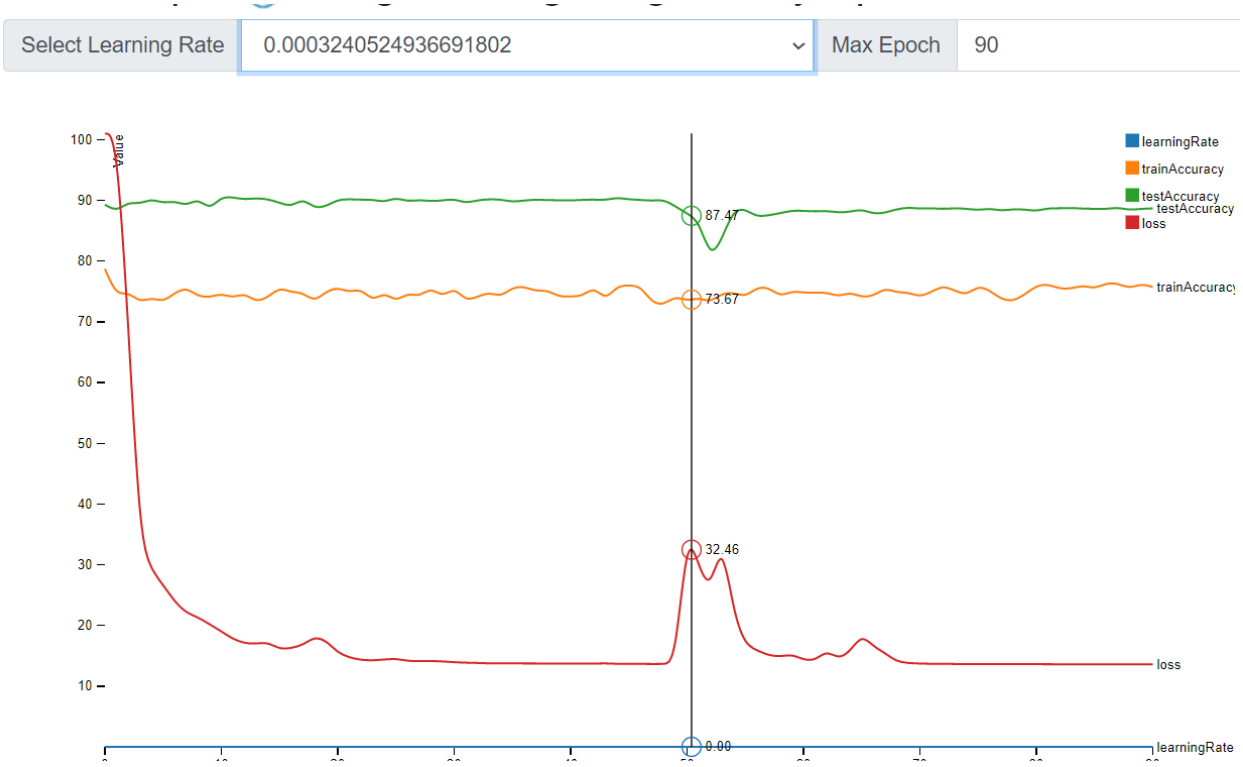
For visualizing the data, I have made a different page with tabs like this-



In the dashboard, this are the red marked elements-

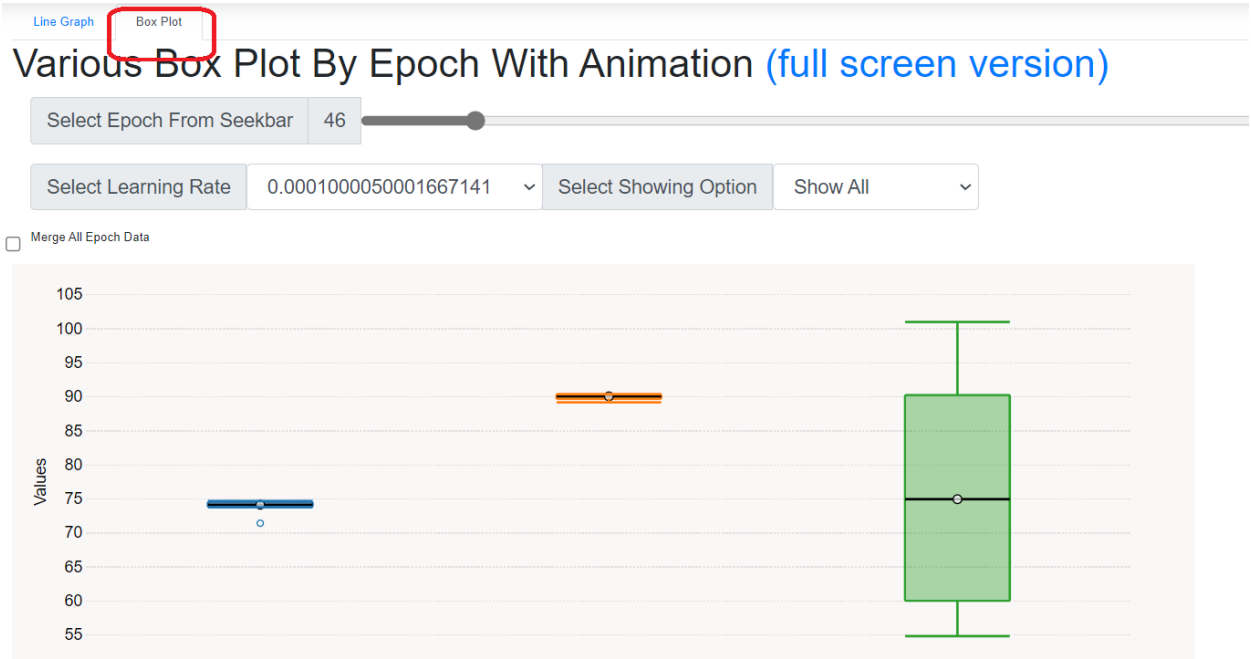
1. Page marker and navigator for each page
2. Tab for switching information
3. Different learning rate selector
4. Maximum epoch selector
5. The markers which representing color of line graph for Learning Rate, Train Accuracy, Test Accuracy and Loss.

There are markers on the graph which is controlled by mouse, which is helpful for visualization, interaction and getting the more accurate value from the graph like this-

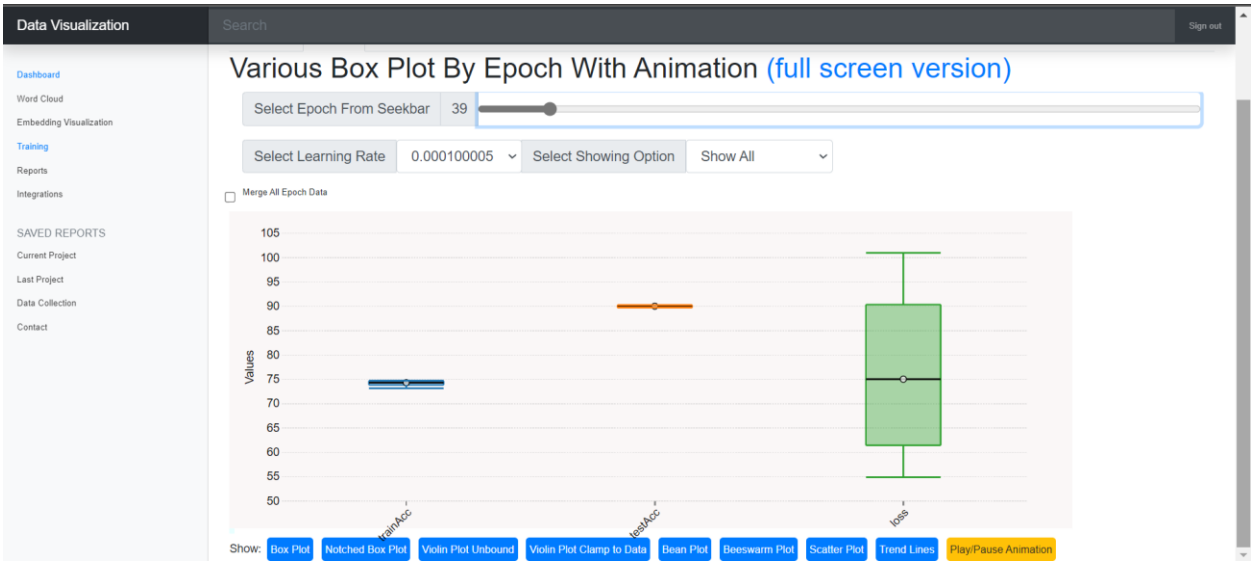


So, by moving the mouse/ hovering the figure by mouse, the values are seen with the values so that we can know what the exact value at that epoch with a specific learning rate is which is a great interactive way to visualize all data.

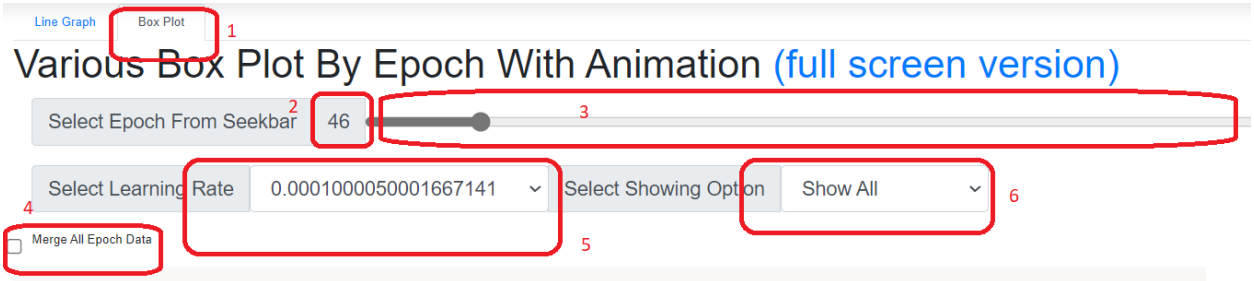
Then there is another tab for visualizing data with different types of box plot-



Dashboard with that tab is looking like this-



There are some other interaction tools which are discussed in here step by step-



In the image, there are markers of different controllers which are discussed in here-

1. Tab switcher

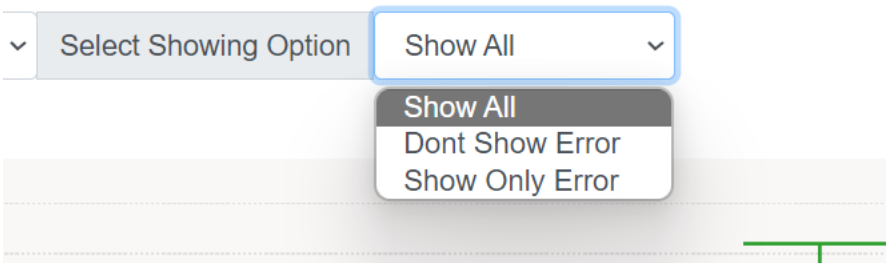
2. Epoch value displayer

3. Epoch selector seek bar

4. Merge all epoch data button

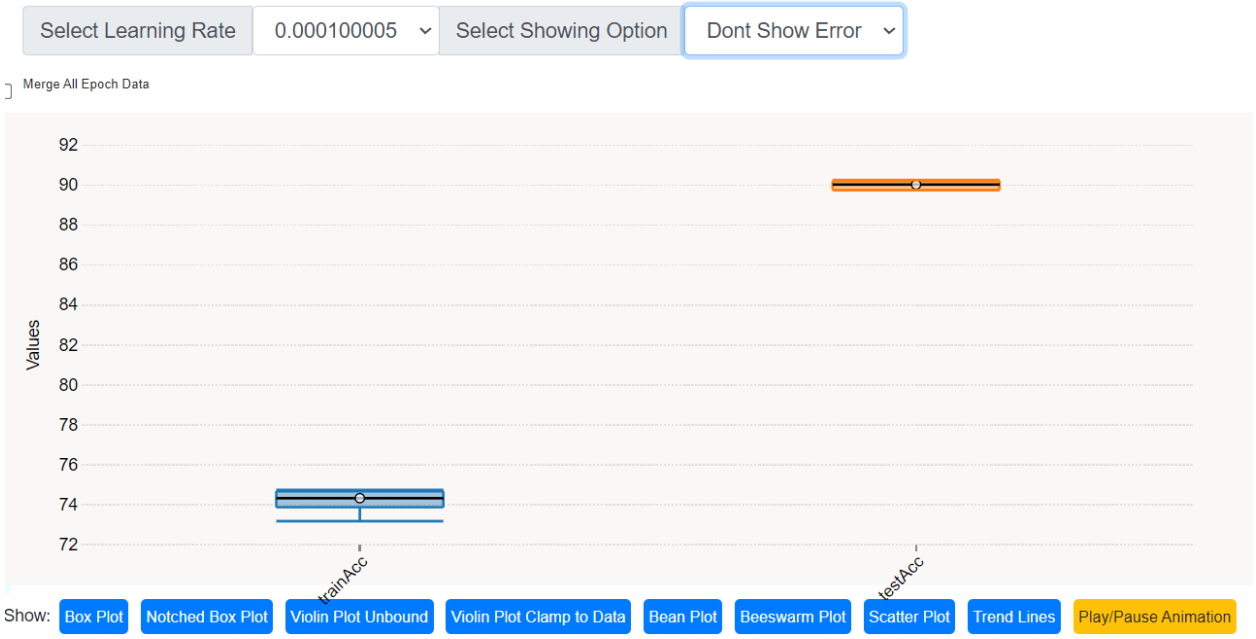
5. learning rate selector

6. option selector where other options are like this-



Where show all is showing training accuracy, test accuracy and loss and other options are like that.

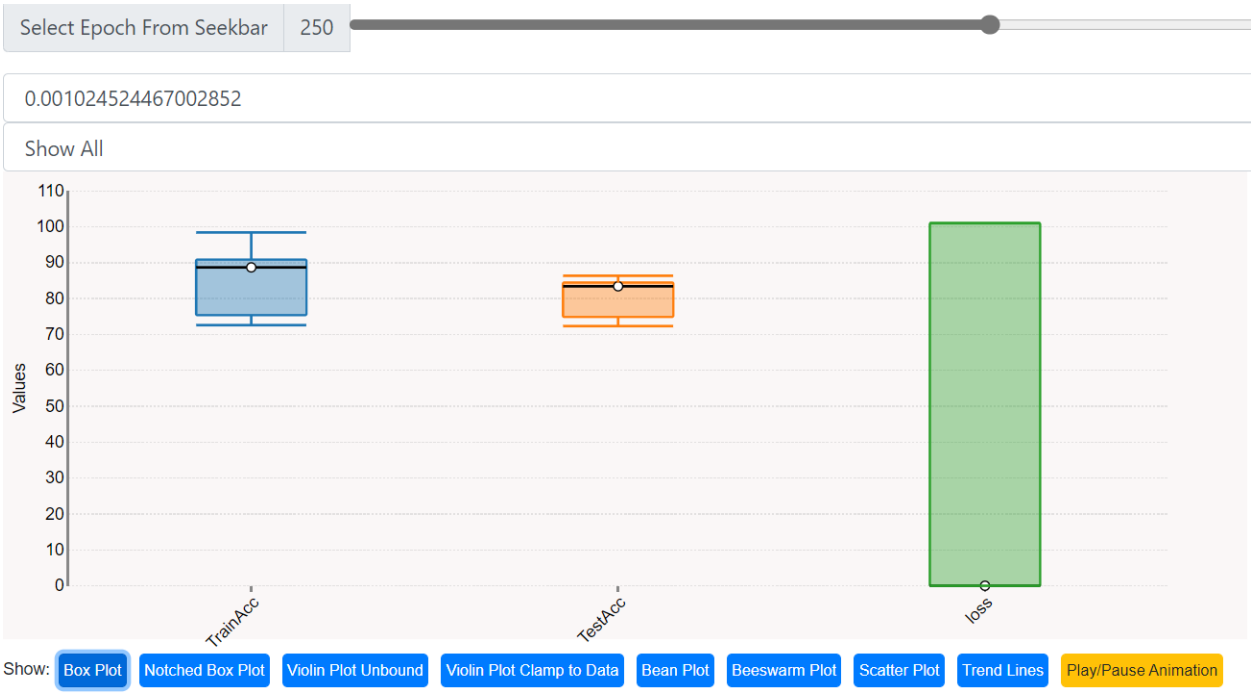
So, if we select “Don’t Show Error” then it is looking like this-



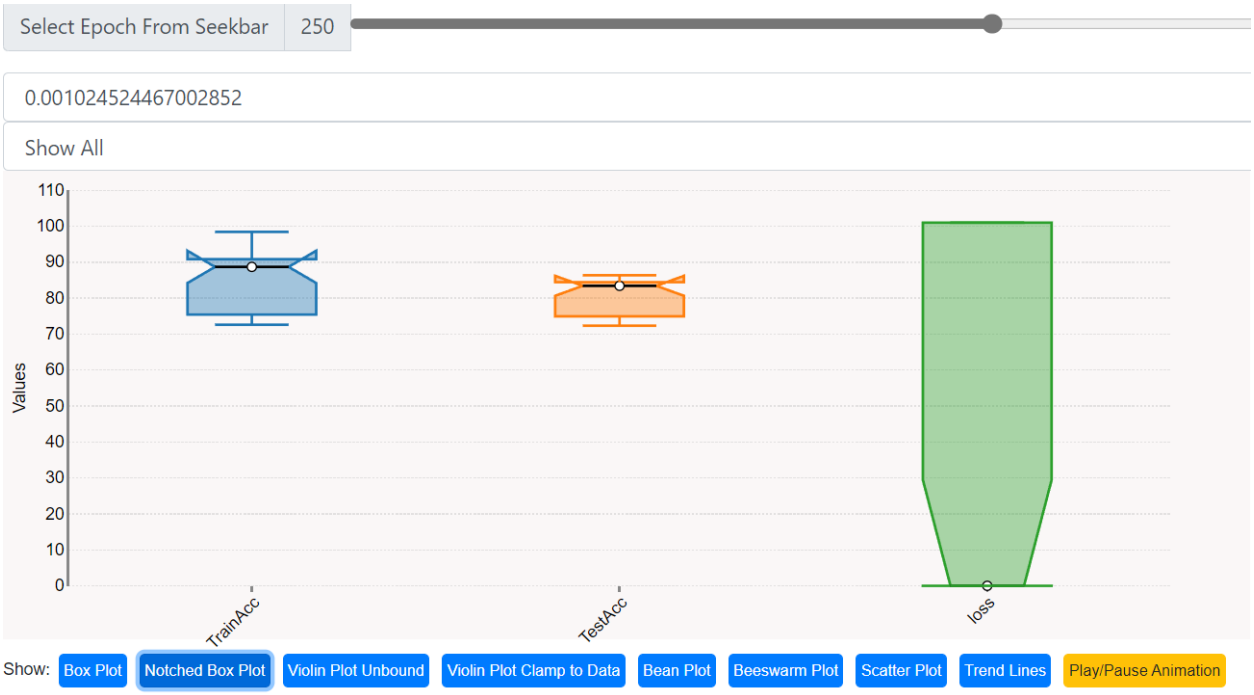
And show only error is also like that.

There are 9 options on bottom, which I am going 1 by 1-

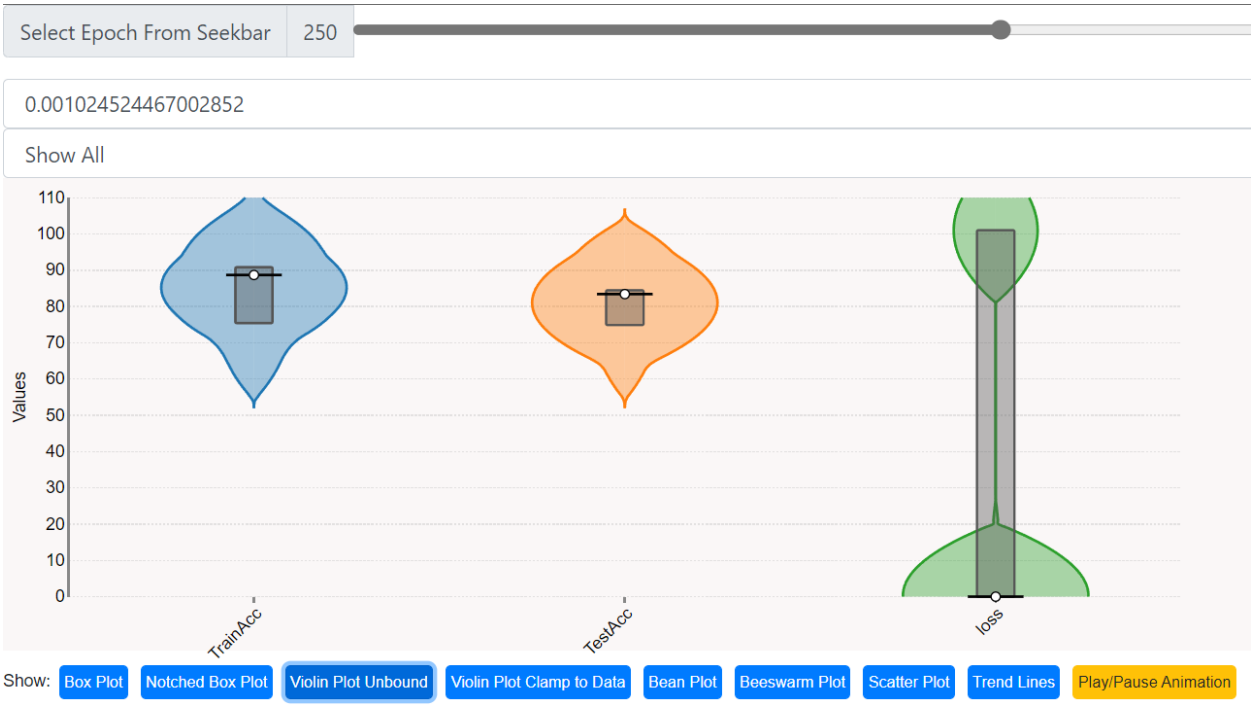
1. Box Plot



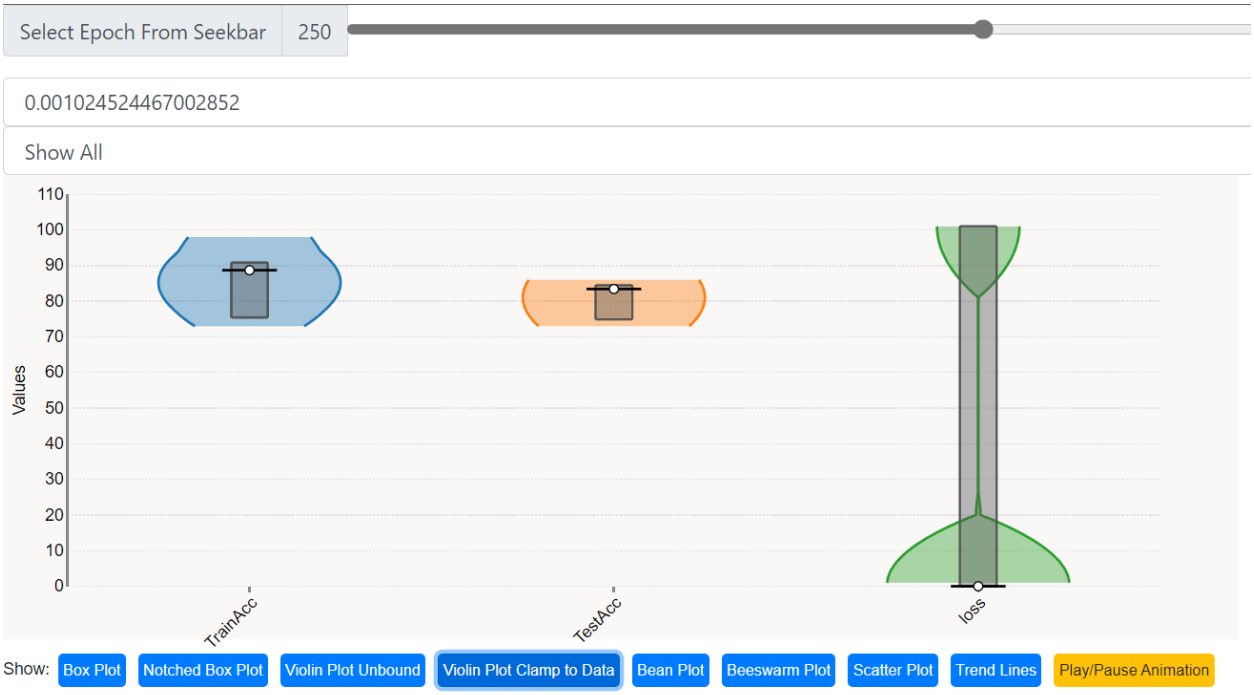
2. Notched Box Plot



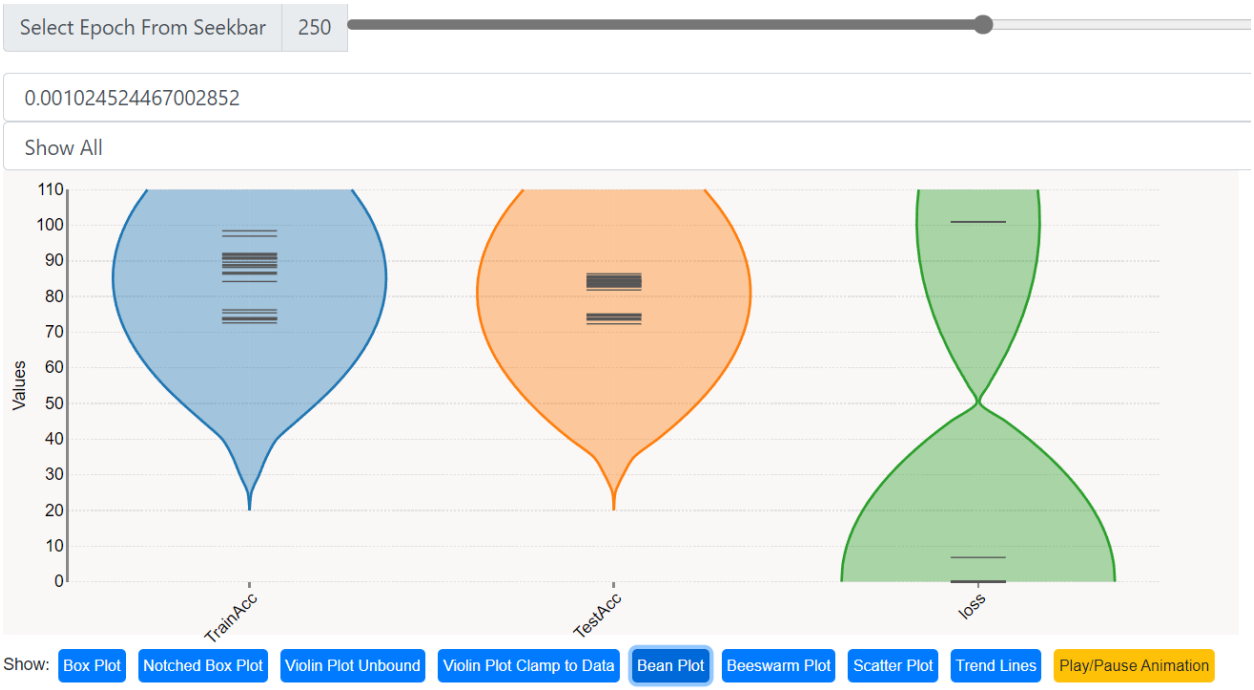
3. Violin Plot Unbound



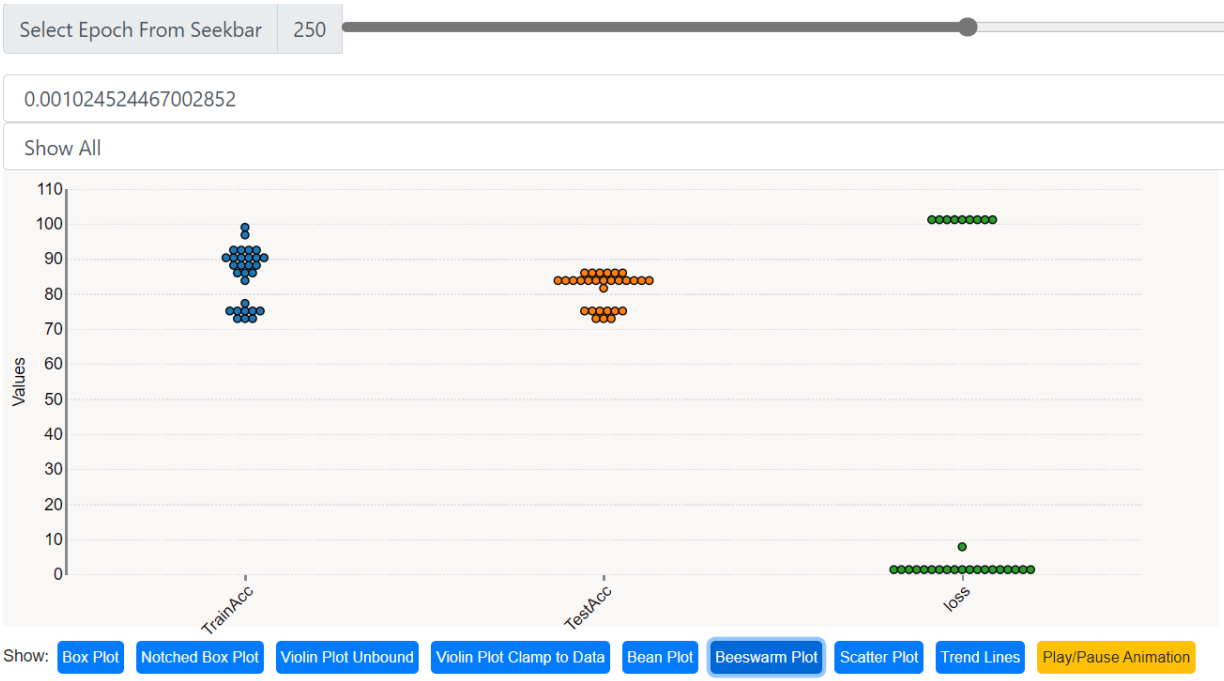
4. Violin Plot Clamp to Data



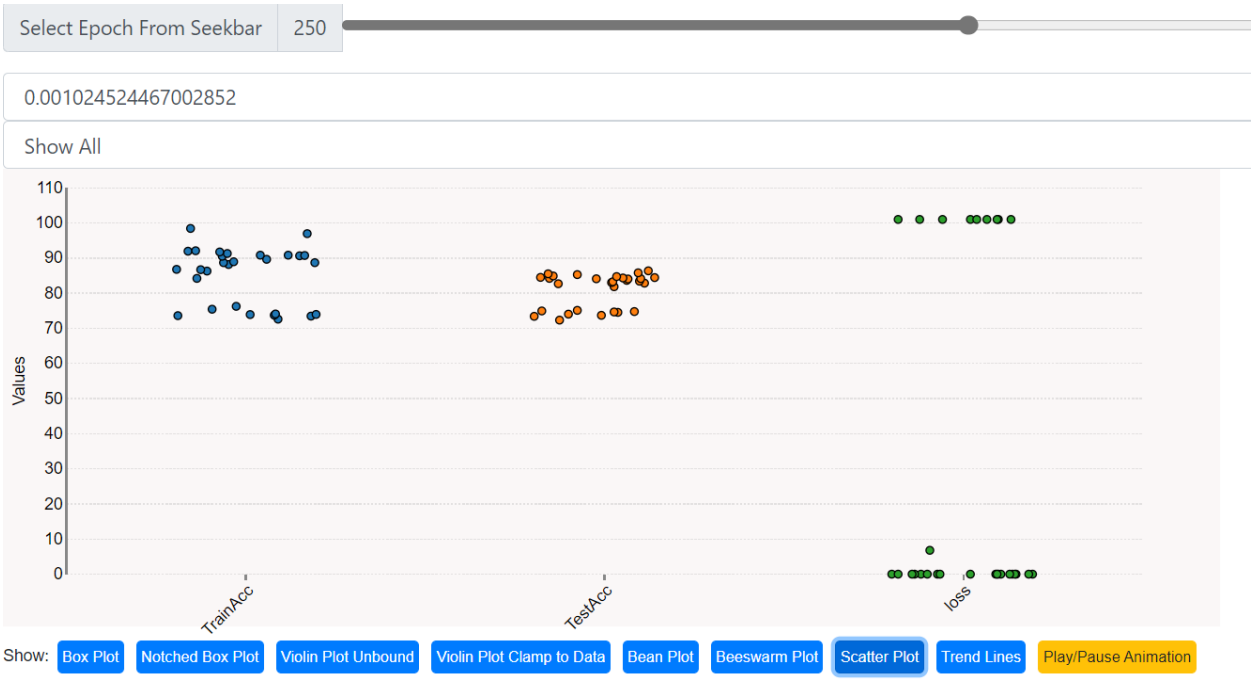
5. Bean Plot



6. Beeswarm Plot

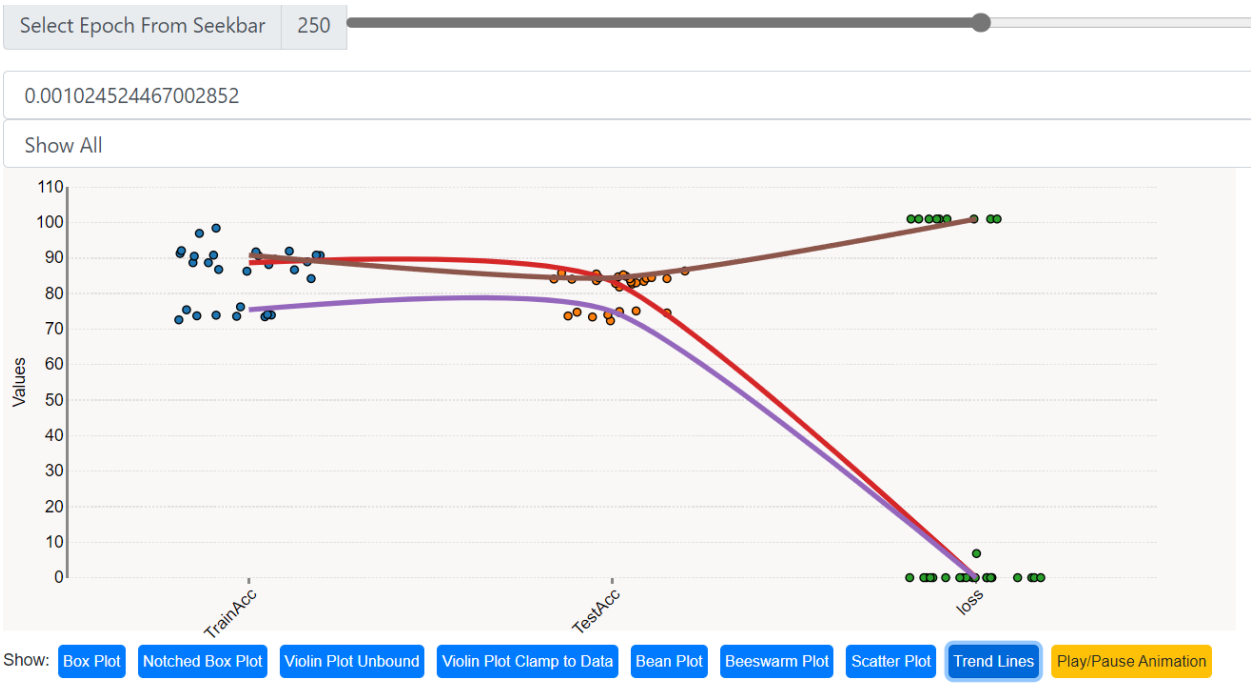


7. Scatter Plot



8. Trend Lines

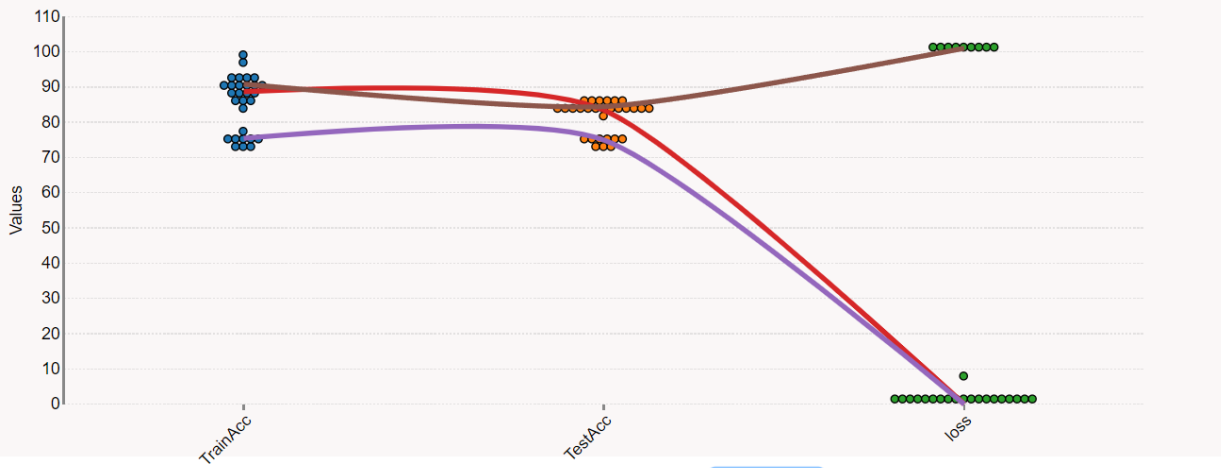
Can be used with different plots like this



Select Epoch From Seekbar 250

0.0010245244670028 52

Show All

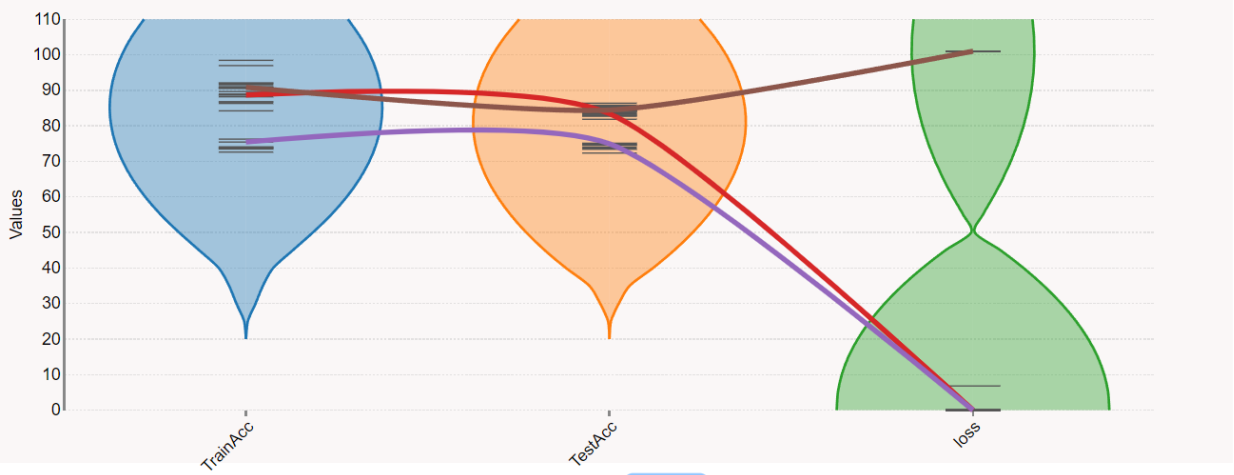


Show: Box Plot | Notched Box Plot | Violin Plot Unbound | Violin Plot Clamp to Data | Bean Plot | Beeswarm Plot | Scatter Plot | Trend Lines | Play/Pause Animation

Select Epoch From Seekbar 250

0.0010245244670028 52

Show All

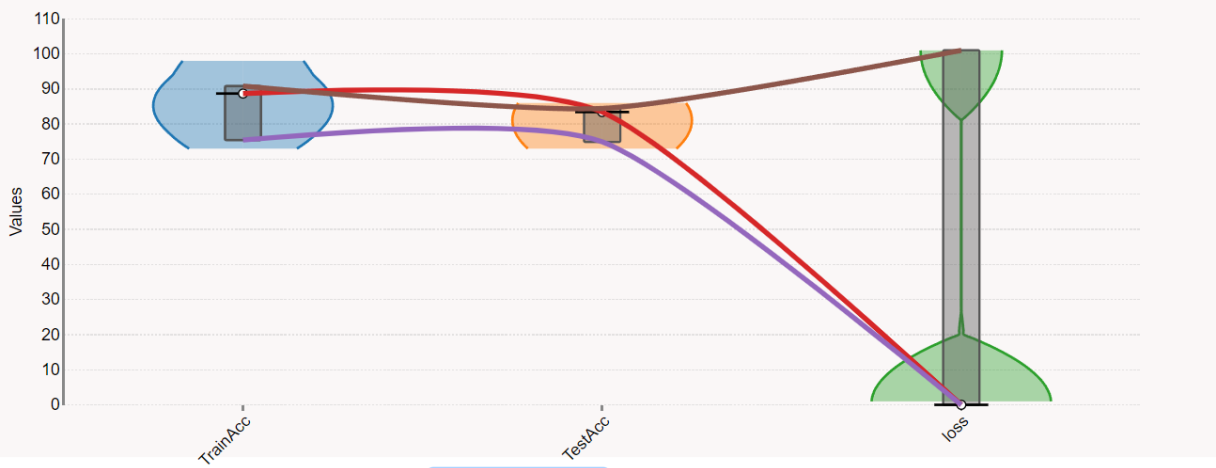


Show: Box Plot | Notched Box Plot | Violin Plot Unbound | Violin Plot Clamp to Data | Bean Plot | Beeswarm Plot | Scatter Plot | Trend Lines | Play/Pause Animation

Select Epoch From Seekbar 250

0.0010245244670028 52

Show All

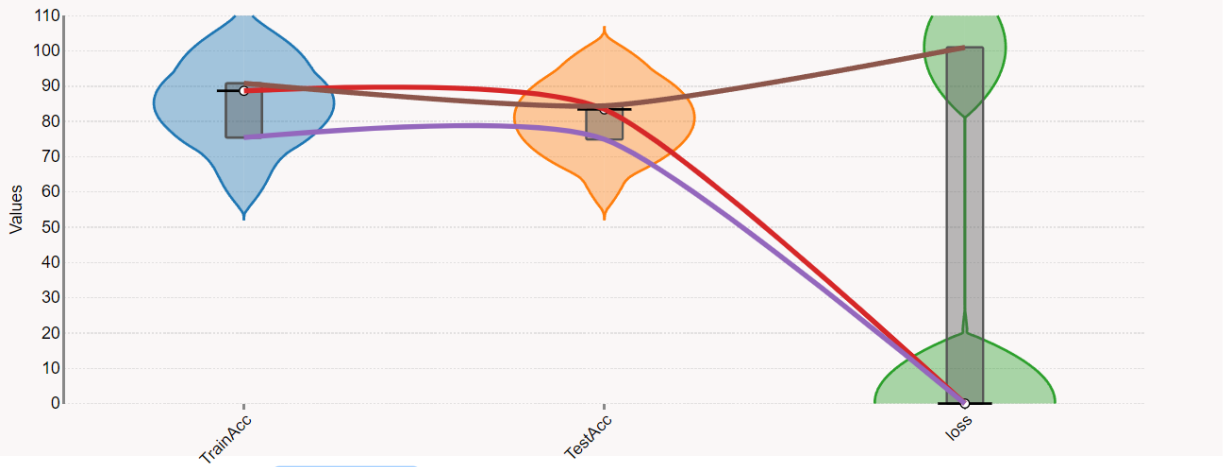


Show: Box Plot | Notched Box Plot | Violin Plot Unbound | Violin Plot Clamp to Data | Bean Plot | Beeswarm Plot | Scatter Plot | Trend Lines | Play/Pause Animation

Select Epoch From Seekbar 250

0.0010245244670028 52

Show All

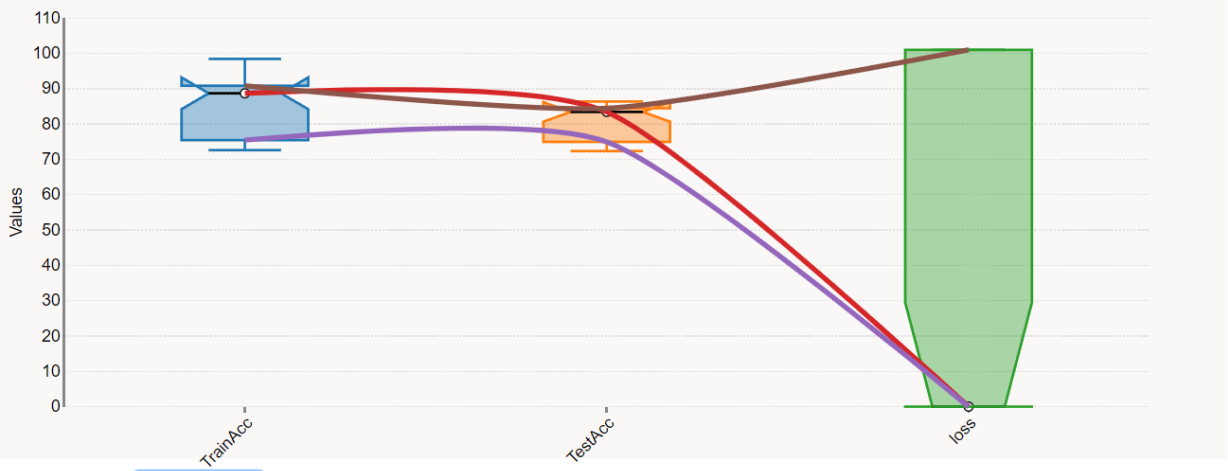


Show: Box Plot | Notched Box Plot | Violin Plot Unbound | Violin Plot Clamp to Data | Bean Plot | Beeswarm Plot | Scatter Plot | Trend Lines | Play/Pause Animation

Select Epoch From Seekbar 250

0.0010245244670028 52

Show All

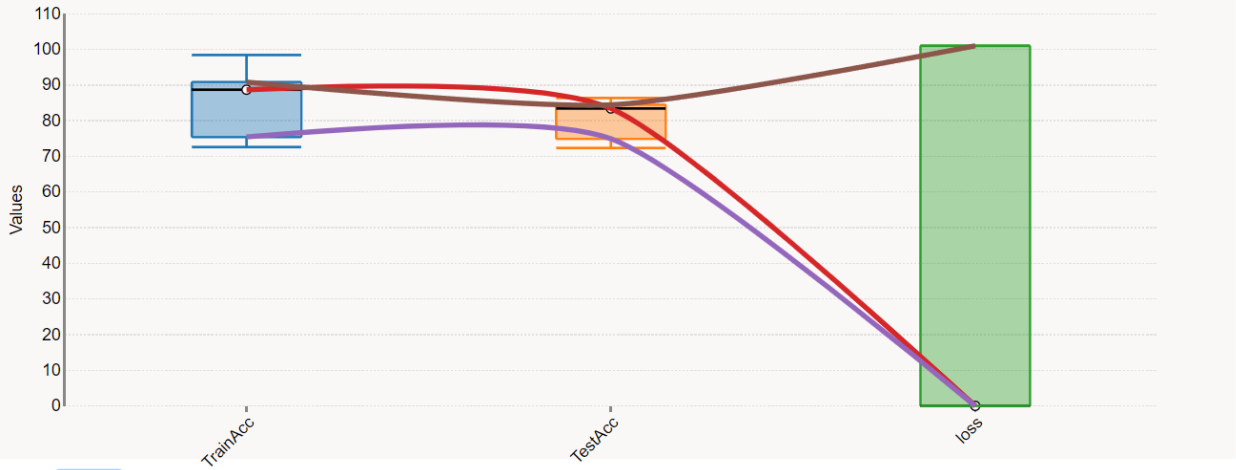


Show: Box Plot | Notched Box Plot | Violin Plot Unbound | Violin Plot Clamp to Data | Bean Plot | Beeswarm Plot | Scatter Plot | Trend Lines | Play/Pause Animation

Select Epoch From Seekbar 250

0.0010245244670028 52

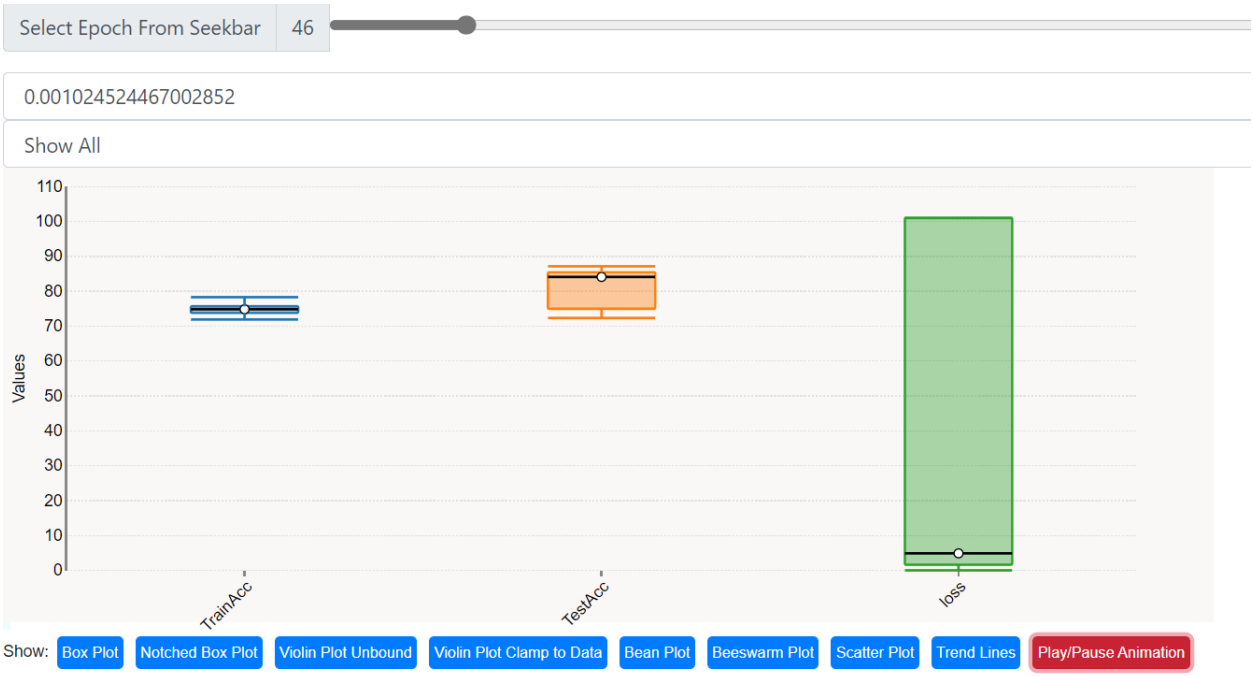
Show All



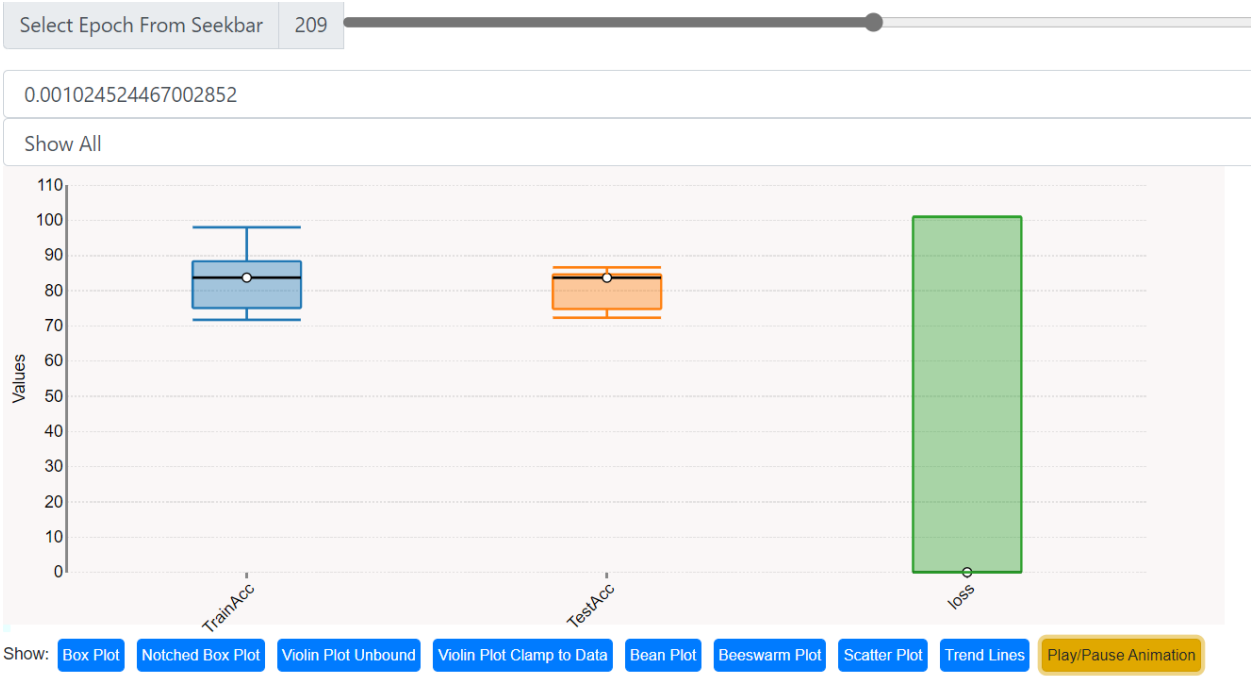
Show: Box Plot | Notched Box Plot | Violin Plot Unbound | Violin Plot Clamp to Data | Bean Plot | Beeswarm Plot | Scatter Plot | Trend Lines | Play/Pause Animation

9. Play/Pause Animation

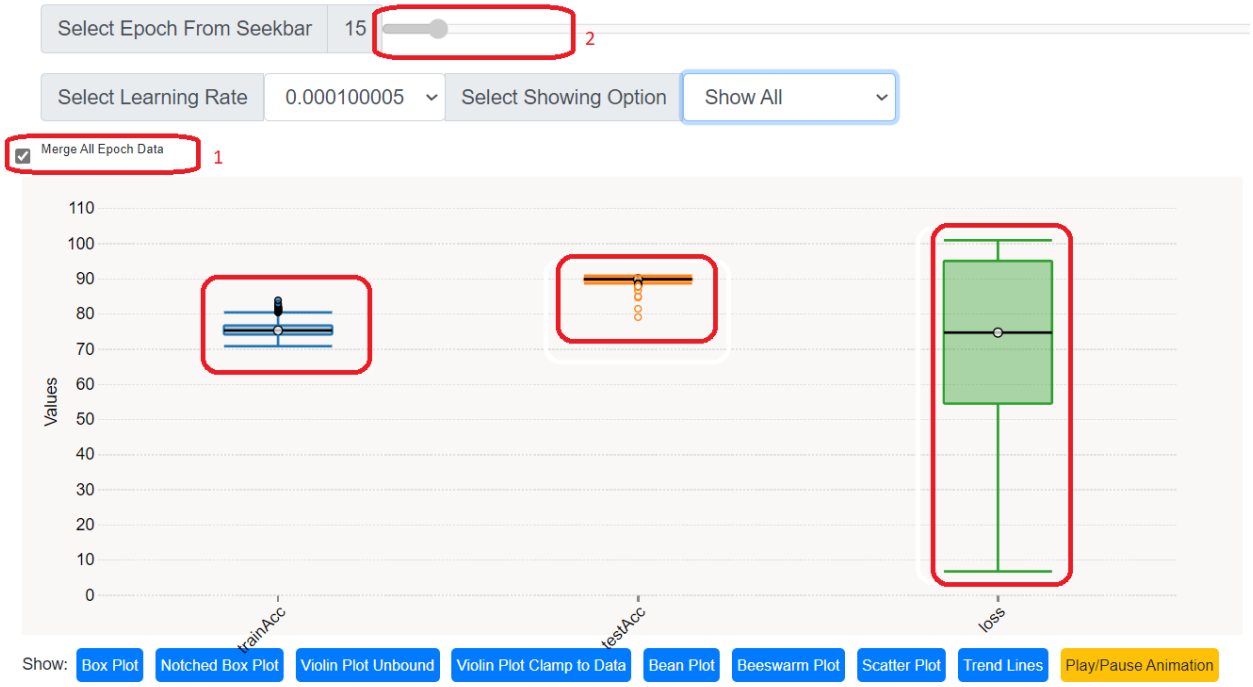
Play and pause animation with the button and after clicking, the color should be changed depending on status-



Red button means animation is on and if we stop by clicking, then it is looking like this-

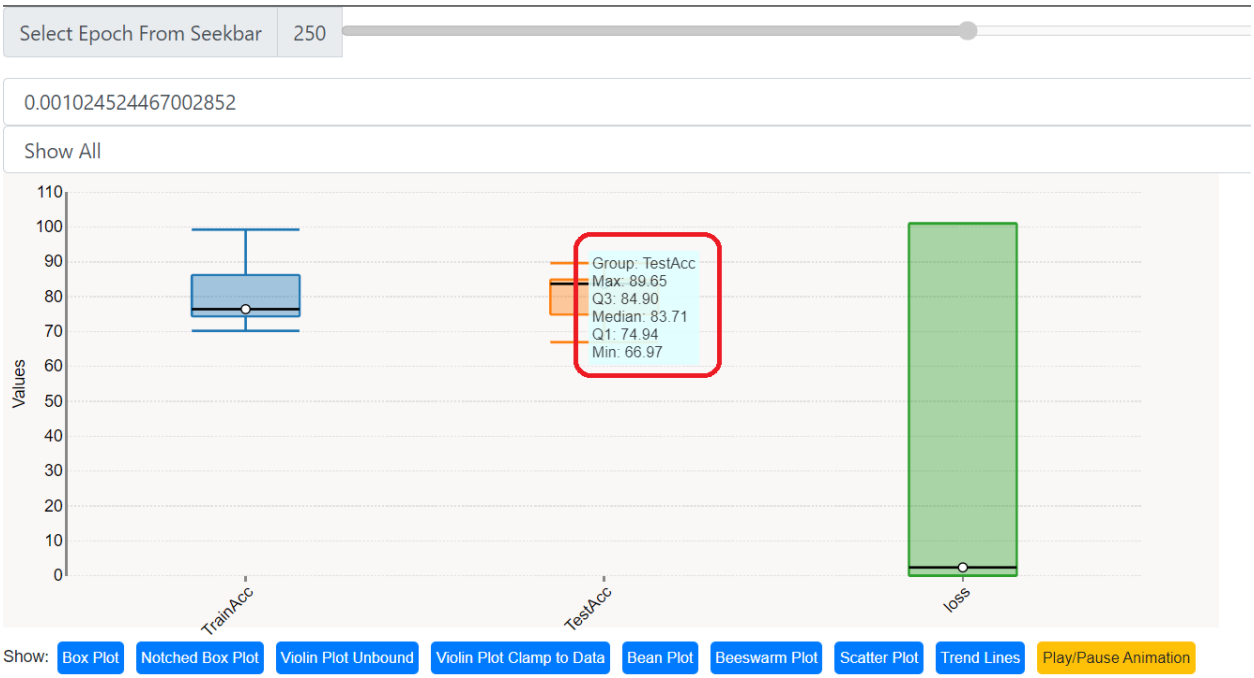


10. Show All



If show all button (marked on 1) is pressed, the seek bar is becoming disabled and vise versa and all data are merged and showed on the different kind of plots.

And on mouseover, the plot details can be seen like this-



On the plot, we are seeing that train accuracy has a range of 70-100% and mean value is around 75%.

For test accuracy, Max and min value is 89.65 and 66.97, Q3 is 84.90, Q1 is 74.94 and Median is 83.71.

For loss, median is close to 0.

Result Analysis-

All these visualizations are for the selected learning rate. And all the interactions can be combined which are shown in here. With these tools, I can easily select a good learning rate where I found **0.001156668** is the best learning rate and **103** would be good epoch for the learning rate for binary classifier training.

-X-