Comparing and Investigating HD* data using Dimensionality Reduction Algorithms

Tushita Gupta, Tanmay Shravge
gupta.tush@northeastern.edu; shravge.t@northeastern.edu

Problem Definition

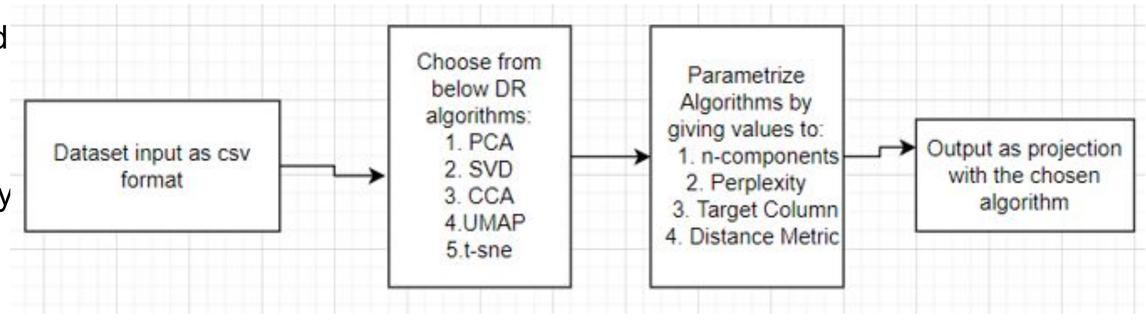
- Given a high dimensional dataset
- Want a dimensionality reduction tool that:
  - Transforms the data from a high-dimensionality into a low dimensionality space
  - Compares different dimensionality reduction algorithms
  - Assists in analyzing the algorithm which fits best by outputting the plots of DR* algorithms

Existing Methods

- Website: helps to understand the parameters of the t-SNE algorithm
  How to use t-SNE effectively. Distill, 2016[M. Wattenberg, F. Viegas, and I. Johnson]
- Probing Projection: Interaction techniques for interpreting arrangements and errors of dimensionality reduction algorithms
  Probing Projections [J. Stahnke, M. Dork, B. Muller, and A.Thom]
- Distiller tool: Provides users with advice on selecting a decent DR algorithm
  Dimstiller [Ingram, T.Munzner, V.Irvine, M.Tory, S.Bergner, and T.Moller]

Proposed Method

- Dimensionality Reduction Assist: A visual analysis tool to visualize and compare different dimensionality reduction algorithms
- Used PCA, SVD, CCA, UMAP, t-sne for Dimensionality reduction
- The workflow can be seen in the figure on the right side.
- Performed two use cases to compare how different dimensionality reduction algorithms work on text and image datasets
- Evaluated the results by visualizing the output using plots

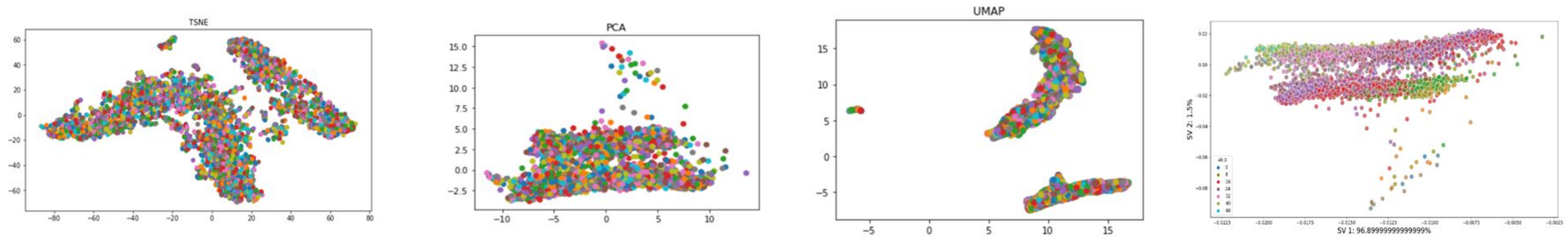


Data Description & Experimental Setup

- Textual Data: Satellite Dataset taken from unsupervised anomaly detection
  - Consisting of 5099 rows and 37 columns
  - This is the Use Case 1
- Image Data: Sign Language Dataset consisting of cropped image montage panel of various users and backgrounds for American Sign Language letters
  - Consisting of 27455 rows and 785 columns (pixels)
  - This is the Use Case 2

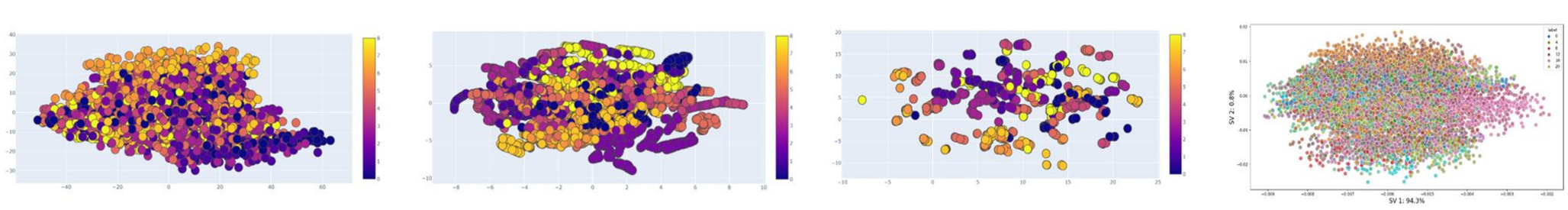
Results

- Use Case 1
  - Deduced from the results in figure below, that in linear methods PCA and SVD have almost similar results in categorizing the data while in non-linear methods UMAP performed better than t-sne.
  - Observe that UMAP performs very well on the dataset.



TSNE PCA UMAP SVD

- Use Case 2
  - It can be observed from figure below, PCA and SVD did not work quite well in categorizing the different signs. TSNE managed to do better work on separating the clusters. However, it took a very long time to compute its embeddings.
  - Observed UMAP performs very well on the dataset, the most effective manifold learning in terms of displaying the different clusters with clear separations.



PCA TSNE UMAP SVD

Discussion of Results

- Dimensionality reduction algorithms results vary when applied on different datasets
- Non-linear methods like t-sne and UMAP provides better visualization and categorization the data as compared to Linear methods(SVD,PCA)
- UMAP performed very well in categorizing both image and text data

Takeaway Points & Future Work

- As a first step in demonstrating how visually comparing DR results might help people understand how different DR algorithms behave
- There are numerous opportunities for future employment. One intriguing suggestion is to improve the tool's robustness by adding more dimensionality reduction techniques, and parameters
- Improve the tool's front end to make it more aesthetically pleasing.

* HD = High-Dimensional data; DR = Dimensionality Reduction