International Refereed Journal of Engineering and Science (IRJES)
ISSN (Online) 2319-183X, (Print) 2319-1821
Volume 13, Issue 1 (Jan.-Feb. 2024), PP. 22-27

Enhancing Semantic Understanding by Visualizing Sentence-Level Embeddings with UMAP

Akshata Upadhye
** (Data Scientist, Randstad USA)*

Abstract: *The field of Natural Language Processing and Machine Learning are advancing rapidly. Due to these advances, various new architectures to train the language models and various new language models are introduced very frequently. These language models can be used in various applications involving text data. Since the number of choices available are high it is very important to have the right tools to evaluate these language models and in such a scenario visualization can help the researchers understand the semantic relationships within data and it can also be used to evaluate if the language model used to extract the features from the text data is able to model these semantic relationships. Since text data is typically high dimensional it is necessary to use dimensionality reduction techniques to be able to visualize the text data. Therefore, in this paper we have discussed various dimensionality reduction techniques and have demonstrated how UMAP can be used for dimensionality reduction to visualize sentence level embeddings.*

Keywords*: articles, UMAP, sentence, embeddings, dimensionality.*

Date of Submission: 01-01-2024 Date of acceptance: 10-01-2024

I. INTRODUCTION

Visualizing sentence embeddings is becoming increasingly important in the field of Natural Language Processing(NLP), as the applications of machine learning and NLP continue to evolve. One of the preliminary challenges when working with text data is the high dimensionality. Sentence embeddings are high-dimensional vectors used to represent text sentences semantically in an n-dimensional space. Visualizing the sentence embeddings can be helpful to the researchers and developers dealing with the text sentences to understand how the model encodes the text information while preserving the semantic relationships.

Additionally, visualizing sentence embeddings can have various other useful applications. Visualizing sentence embeddings can be useful to evaluate the quality of a language model used to extract these embeddings. If the group of dissimilar sentences that are clustered together might be an indicator that the language model requires more fine tuning. Therefore, visualizing can aid in debugging complex language models to improve their performance. Similarly, the visualization can be helpful in comparing sentence embeddings extracted using various language models in order to select the best suitable embedding representation for a particular task or to benchmark the performance of various language models.

With the recent developments in the Large Language Models(LLM's) it can be very helpful to visualize sentence embeddings extracted from LLM's to understand how these models capture semantic relationships and group the similar sentences together. Specifically with the rise in the artificial intelligence generated content, visualizing sentence embeddings can be helpful in detecting bias. For instance, if a group of sentences are grouped together and are always further apart, then this is an indicator of bias in training data.

In addition to all this research and developer applications, visualizing sentence embeddings can be used to educate a wide group of audience especially when trying to convey concepts and findings in NLP in an intuitive manner. Therefore, our goal is to design a system to extract meaningful embedding representation of text data and to visualize the same by using dimensionality reduction technique known as UMAP.

II. RELATED WORK

Dimensionality reduction techniques are crucial when it comes to visualizing the high-dimensional representations of text data also known as sentence embeddings. These techniques are useful to obtain lower-dimensional approximation of high-dimensional data typically in 3D or 2D spaces. Here is a summary of some commonly used dimensionality reduction techniques for text embedding visualization:

2.1 PCA: Principal Component Analysis

PCA is a popularly used linear dimensionality reduction technique and it can be applied to text embeddings. PCA projects the data into a lower-dimensional space by identifying the principal components of the data while maximizing variance. This method is computationally efficient and can be useful to get insights

into the most significant dimensions of the data. One of the drawbacks of this technique is that it may not capture the nonlinear relationships as effectively as methods such as t-SNE or UMAP [1].

2.2 MDS: Multidimensional Scaling

Multidimensional Scaling is a dimensionality reduction technique and has two variants known as classical and nonmetric MDS. MDS preserves the pairwise distances or similarities between data points in the lower-dimensional space. Therefore, it can be effective when the relationships in the data are primarily distance-based [2].

2.3 t-SNE: t-Distributed Stochastic Neighbor Embedding

t-Distributed Stochastic Neighbor Embedding is a technique used for nonlinear dimensionality reduction and it is effective for visualizing high-dimensional text embeddings. t-sne captures the complex relationships and clusters in the data which makes it suitable for exploring and visualizing text data specifically word embeddings or sentence embeddings. t-sne for dimensionality reduction requires tuning the perplexity parameter for getting optimal results [3].

2.4 UMAP: Uniform Manifold Approximation and Projection

UMAP is a nonlinear dimensionality reduction technique which falls under the manifold learning category. UMAP aims to project high-dimensional data into a lower-dimensional manifold while preserving important relationships between the data points. UMAP is popularly used in the field of genetics [4]. UMAP has several advantages over t-sne including its ability to efficiently capture both local and global structures within the data. Therefore, it is useful for preserving semantic relationships in word or sentence embeddings [5]. Due to all these advantages until now UMAP has been popularly used in the field of biology in applications such as visualizing single-cell data [6], physical and genetic interactions [7], population genetics [8], population structure and phenotype heterogeneity in large genomic cohorts [9], etc. In this paper we are demonstrating how it can be leveraged to visualize text data.

III. BACKGROUND

3.1 Text representation using Doc2vec

To demonstrate how UMAP can be used to visualize sentence embeddings, we selected the Doc2vec neural network-based algorithm to learn distributed representation of sentences in the embedding space [10]. In the original paper there are two different frameworks that can be used to train the Doc2Vec model:

- **Paragraph Vector-DBOW (Distributed Bag of Words):** While training the model with this approach the objective is to predict the next word in a paragraph by using the context of word vectors.
- **Paragraph Vector-DM (Distributed Memory):** In this approach an additional paragraph vector is concatenated or averaged with the context of words which then is used to predict the next word in the training process. The Doc2vec model will be helpful to extract sentence level embeddings for the documents in our dataset.

3.2 Dimensionality Reduction using UMAP

UMAP is a graph-based dimensionality reduction technique categorized under the manifold learning and is specifically used for non-linear dimensionality reduction. It offers various advantages compare to t-SNE because it preserves both the local and global structures within the data due to which the dissimilar data points will be grouped farther away and the similar data points will be closer, and the distances will be preserved meaningfully such that the global structure is maintained in the lower dimensions. Additionally, UMAP is computationally faster as compared to t-SNE.

The main hyperparameters in the UMAP are the nearest neighbors, minimum distance, and the number of dimensions used for projection. The nearest neighbors parameter is useful in preserving the overall structure of the data. The lower values of nearest neighbors will result in local structure being preserved more than the global structure whereas higher values of nearest will help preserve the global structure. The minimum distance parameter is used to preserve the effective minimum distance between the data points. Smaller values of the minimum distance keep the points closely connected or clustered together and the larger values will keep them at a distance. The number of dimensions is our expected number of dimensions. Due to the advantages and the parameter tuning flexibility that the UMAP offers, it can be well suited to visualize text data.

IV. METHODOLOGY

In this section we discuss how we generate sentence level embedding, implement dimensionality reduction, and use it for visualization.

4.1 Phase 1: Preprocessing documents

In order to use the text data from the documents dataset *D* every document *Di* goes through several preprocessing steps. The text data from every document *Di* is tokenized, then the stop words are removed, and every token is lemmatized and the frequently occurring n-grams are added to the list of tokens.

4.2 Phase 2: Generating sentence level embeddings

In this stage a subset of preprocessed documents is used to train a Doc2vec model implemented in the gensim library using the PV-DM architecture. Then the Doc2vec embedding model is used to extract a sentence level 20-dimensional embedding representation of the document *Di* for all the documents in the entire document collection *D*.

4.3 Phase 3: Training UMAP model for dimensionality reduction

Once we have extracted the embedding representation which is in 20 dimensions, we will then train a UMAP model to generate the 2-dimensional representation of the 20-dimensional sentence embeddings.

4.4 Phase 4: Visualization

Finally, we use the 2D embeddings obtained from UMAP and the true class labels to visualize the documents using the matplotlib scatterplot.

V. DATASETS

For the purpose of visualizing the sentence level embeddings we collected 750 research papers from 6 categories - Chemistry, Data Mining, Wireless Sensor Networks, Graph Theory, Statistics and Cyber Security from Elsevier's open access science journals. The class distribution within the dataset is shown in Table 1.

Table 1 Class Distribution in the dataset

Class	Number of documents
Chemistry	125
Data Mining	125
Wireless Sensor Networks	125
Graph Theory	125
Statistics	125
Cyber Security	125

VI. RESULTS AND INTERPRETATION

In this section we will discuss the impact of the crucial hyperparameters while using UMAP for dimensionality reduction through visualization. Each of these parameters influence the visualization of the sentence level embeddings. The figures 1-8 are scatter plot of 2D UMAP embedding representation and we use the true class labels to show documents belonging to the same group by assigning them the same color.

6.1 Nearest Neighbors

The Nearest Neighbors is one of the crucial hyperparameters which helps define the local neighborhood of data points while construction of a low dimensional approximation of data. The number of nearest neighbors value determines how many neighbors should be present in the local neighborhood for each data point. When we pick smaller values for the nearest neighbors in the range 2-10 the resulting low dimensional representation will have fine-grained structure of the local neighborhood as we can see in the figure 1 having nearest neighbors = 3 and figure 2 having nearest neighbors = 8. Whereas when we pick larger values in the range 11- 200 for the nearest neighbors, the resulting low dimensional representation will have larger local neighborhoods which helps preserve the global structure within the data, as we can see in figure 3 having nearest neighbors = 20 and figure 4 having nearest neighbors = 200. Therefore, from figures 1-4 we can see the nearest neighbors parameter influences the density of the neighborhood which in turn influences how much importance is given to the local versus the global structure.

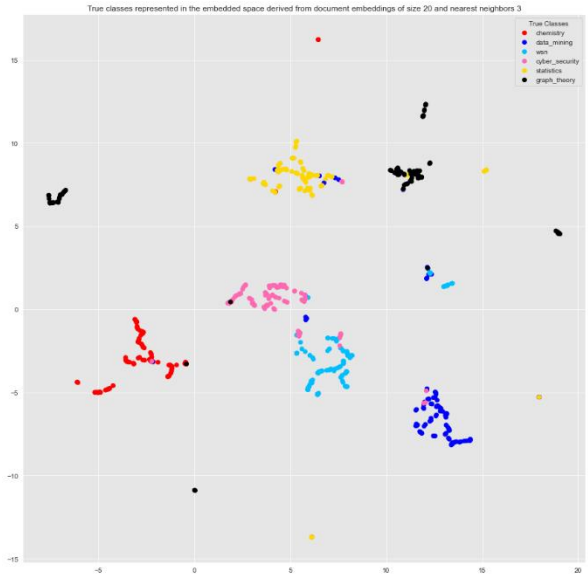

Fig. 1. UMAP with 3 neighbors

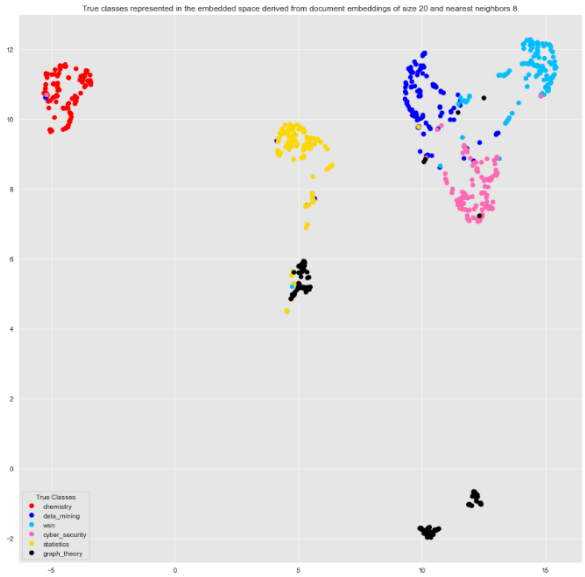

Fig. 2. UMAP with 8 neighbors

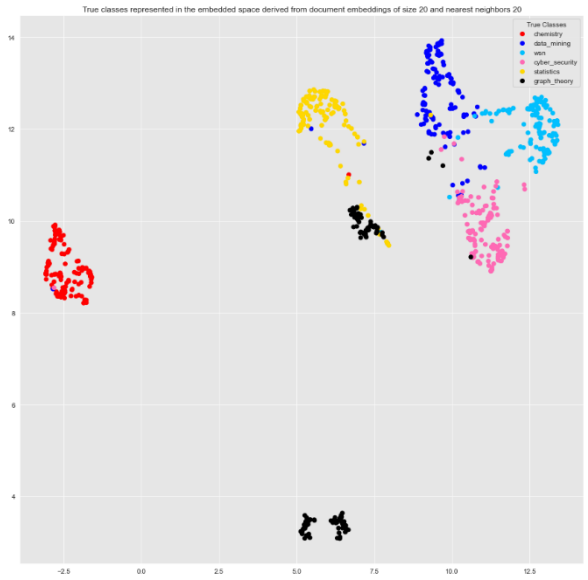

Fig. 3. UMAP with 20 neighbors

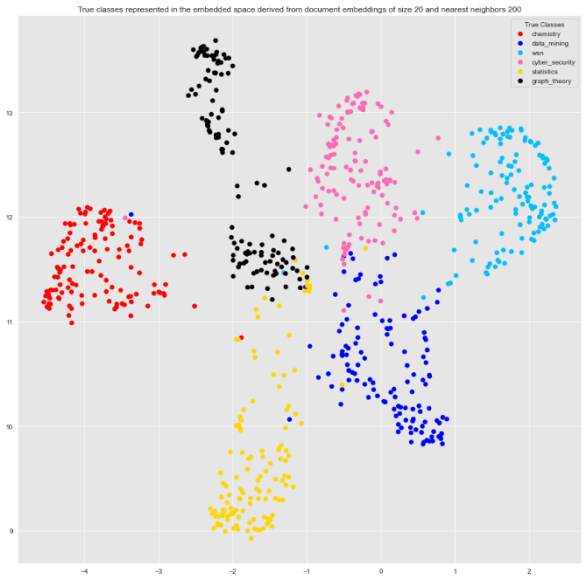

Fig. 4. UMAP with 200 neighbors

6.2 Minimum Distance

The Minimum Distance is another important hyperparameter used in UMAP to set a threshold for how close the set of two points should be in the high dimensional space for them to be neighbors in the low dimensional representation. Lower values of the minimum distance parameter will create tightly connected neighborhoods which can lead to a more detailed representation locally but might result in overlaps between the data points thus leading to a loss of global structure which can be seen in figure 5. As we increase the value of minimum distance threshold, the resulting low dimensional representation will have greater separation between the points which will create a coarser representation of the neighborhood with less overlaps. We can see this in figure 6 with minimum distance = 0.25, 7 with minimum distance = 0.5 and figure 8 with minimum distance = 0.99, where the points are at a distance which can help maintain the global structure. Hence the minimum distance parameter has a significant influence on the low dimensional embedding generated using UMAP since it affects the minimum distance at which points are allowed to be in the low-dimensional representation. Therefore, it affects the balance between preserving local structure and maintaining global structure in the visualization.

Depending on the specific task that we are trying to accomplish it is necessary fine-tune these hyperparameters to get an interpretable and meaningful low dimensional representation of the data. Although UMAP generates a good approximation of the high dimensional data in a low dimensional space, we should be

aware that in this process some distortions may be introduced, and it is necessary to be aware about these limitations and the effects of using various hyperparameter values.

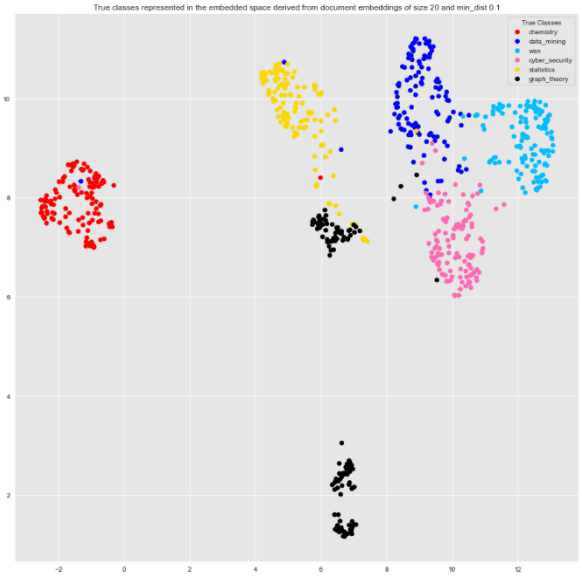

Fig. 5. UMAP with minimum distance = 0.1

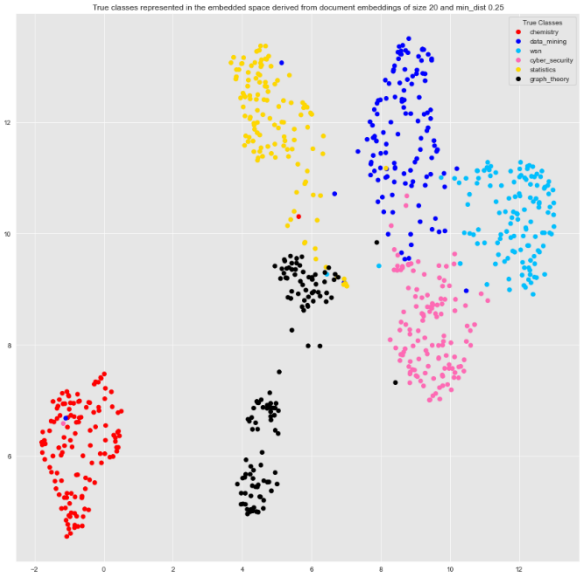

Fig. 6. UMAP with minimum distance = 0.25

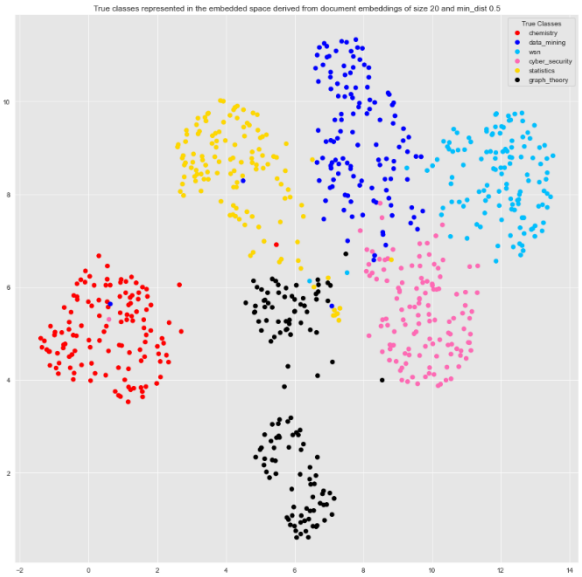

Fig. 7. UMAP with minimum distance = 0.5

Fig. 8. UMAP with minimum distance = 0.99

VII. CONCLUSION

In this paper we have discussed the existing approaches for dimensionality reduction of the high dimensional sentence level embeddings such as PCA, t-SNE, etc. We discussed their advantages and their drawbacks. We also discussed the advantages of using UMAP for dimensionality reduction in the context of text data, in our case sentence embeddings. We have demonstrated by using a step-by-step approach on how to generate sentence embeddings for text data and how to use UMAP for dimensionality reduction and for visualization. Finally, we have also discussed the important parameters that need to be considered while using UMAP and have also highlighted how these hyperparameters could affect the generation of low dimensional embeddings through visualization. Therefore, we have demonstrated that UMAP could be a good tool for visualizing sentence embeddings, and this can be useful in various other tasks such as evaluating text data, evaluating the language models, evaluating the clustering results, and many such applications through visualization.

REFERENCES

[1]. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202

[2]. Ghojogh, B., Ghodsi, A., Karray, F., & Crowley, M. (2020). Multidimensional scaling, sammon mapping, and isomap: Tutorial and survey. *arXiv preprint arXiv:2009.08136.*

[3]. Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

[4]. Di Giovanni, D., Enea, R., Di Micco, V., Benvenuto, A., Curatolo, P., & Emberti Gialloreti, L. (2023). Using machine learning to explore shared genetic pathways and possible endophenotypes in autism spectrum disorder. *Genes*, 14(2), 313.

[5]. McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

[6]. Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W., Ng, L. G., ... & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology*, 37(1), 38-44.

[7]. Dorrity, M. W., Saunders, L. M., Queitsch, C., Fields, S., & Trapnell, C. (2020). Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nature communications*, 11(1), 1537.

[8]. Diaz-Papkovich, A., Anderson-Trocmé, L., & Gravel, S. (2021). A review of UMAP in population genetics. Journal of Human Genetics, 66(1), 85-91.

[9]. Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C., & Gravel, S. (2019). UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS genetics*, 15(11), e1008432.

[10]. Le, Q., & Mikolov, T. (2014, June). Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196). PMLR.