A Comparative Study of Text Document Representation Approaches Using Point Placement-based Visualizations

Hevelyn Sthefany Lima de Carvalho
Departamento de Ciência da Computação
Universidade de Brasília
Brasília, Distrito Federal, Brazil
Email: hev.sthefany@aluno.unb.br

Vinícius R. P. Borges
Departamento de Ciência da Computação
Universidade de Brasília
Brasília, Distrito Federal, Brazil
Email: viniciusrpb@unb.br

***Abstract*—In natural language processing, text representation plays an important role which can affect the performance of language models and machine learning algorithms. Basic vector space models, such as the term frequency-inverse document frequency, became popular approaches to represent text documents. In the last years, approaches based on word embeddings have been proposed to preserve the meaning and semantic relations of words, phrases and texts. In this paper, we focus on studying the influences of different text representations to the quality of the 2D visual spaces (layouts) generated by state-of-art visualizations based on point placement. For that purpose, a visualization-assisted approach is proposed to support users when exploring such representations in classification tasks. Experimental results using two public labeled corpora were conducted to assess the quality of the layouts and to discuss possible relations to the classification performances. The results are promising, indicating that the proposed approach can guide users to understand the relevant patterns of a corpus in each representation.**

I. Introduction

In text classification and analytics, an important concern refers to the representation of text documents. The unstructured nature of original texts demands the use of techniques to transform them to structured manner, so that they can be compared using distance functions, as well as processed by machine learning and visualization techniques. Traditionally, the Bag-of-Words (BoW) and term frequency-inverse document frequency (tf-idf) techniques have been employed in most text mining tasks due to its efficiency to compute the feature vectors, also allowing to determine the similarity between text documents. However, these representations lack to preserve the semantic relationships and the meaning of sentences and words regarding the document context.

The recent advances in modern computers have enabled the development of language models based on deep neural networks which are capable to capture implicit information of texts. In this sense, word embeddings [1] have emerged as powerful approaches to represent words, sentences and documents. However, some visualization techniques, such as the multidimensional projections [2], demands to compute dissimilarities between documents using its underlying feature vectors. This motivated us to investigate the use of word embeddings for document representation as an alternative to the BoW and tf-idf, so that the obtained feature vectors can be successfully employed for text visualization.

We also know that, due to the complex nature of high-dimensional space, any reduction applied by visualization techniques will manifest significant distortions giving misleading results. Therefore, several surveys presented comparisons of techniques to help choose the appropriate method [3]–[5]. However, to the best of our knowledge, there are not many researches focused on combining visualization techniques with word embedding based techniques. This also motivated us to present this more restricted study to better transform and visualize the texts.

This paper describes a comparative study of feature space visualizations using projection techniques and their respective classifications, as well as an investigation of the relationship of bi-dimensional visual spaces' (layout) qualities and its relation to classification performance. For that purpose, we propose a method constituted by text preprocessing, feature extraction using classical approaches and word embeddings, and text visualization based on point placement strategies: Principal Component Analysis (PCA) [6], Isometric Feature Mapping (Isomap) [7], Uniform Manifold Approximation and Projection (UMAP) [8], Locally Linear Embedding (LLE) [9] and t-Distributed Stochastic Neighbor Embedding (t-SNE) [10]. Finally, the assessment of layout quality is performed using metrics regarding neighborhood preservation and cluster separation.

The paper's main contributions are:

- a comparative study of different text representation approaches and how they affect the quality of 2D visual spaces (layouts) generated by visualizations based on point placement;
- a visual approach of feature spaces that allows users to analyze the essential information and meanings of text documents according to the various structured representations;
- a strategy to assess the quality of point placement-based visualization by using well-known metrics and attempting

to make relations with the performance of classification tasks.

This paper is organized as follows. Section II presents some related works on feature engineering and visualization based on multidimensional projections using text corpora and images. Section III details the proposed method and its constituting steps. Section IV describes the experiments to validate the proposed method and discusses the results. Finally, Section V concludes the paper and introduces possibilities for future work.

II. Related Work

In the last decade, visualization approaches have been employed in several natural language processing tasks such as text classification [11], [12], topic modeling [13] and sentiment analysis [14], [15]. Recent researches related to the visual exploration of feature spaces of textual data have also been proposed and are discussed next.

Motta et al. [16] introduce measurements of visual properties and preservation of original space properties and discusses the local and global behavior of projection techniques, considering various mappings of real and artificial datasets. In this way, the study presents strategies for interpreting the layouts, while comparing them regarding some graph-based measures and properties.

Embedding Projector [17] is a web application tool launched by Google as part of the Tensor-Flow framework, for interactive visualization and analysis of high-dimensional data. This can be useful for viewing, examining, and understanding its embedding layers. Currently, the Embedding Projector offers three methods for reducing data dimensionality in visualization processes: Uniform Manifold Approximation and Projection (UMAP), Principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) and custom linear projections.

Shusen Liu et al. [18], on the other hand, present novel approaches to compute linear embeddings of semantic relationships and two novel views for a comprehensive study of analogy relationships. In addition, t-SNE embeddings have been augmented, incorporating per-word distortion metrics, as well as an interactive display of neighboring words in high-dimensional space. In this sense, it was possible to intuitively illustrate the most relevant and reliable features in data.

Particularly, the most of literature researches related to visual text analytics represent text documents using the well-known vector space model, specifically BoW or tf-idf approaches. However, recent techniques based on word embeddings and transformers have shown to be powerful in Natural Language Processing (NLP) tasks in which the meaning and semantic relationships are relevant for the underlying tasks. Therefore, we propose to explore and evaluate the quality of layouts obtained by point placement-based visualizations using corpora presenting different text representations and relate it to the classification performances.

III. Proposed Method

The proposed method is depicted in Figure 1 and receives a corpus (text collection) as input. First, a text preprocessing is performed prior to the feature extraction which can comprises tf-idf [19] and word embeddings techniques: word2vec [1] [20], Global Vectors (GloVe) [21] and Bidirectional Encoder Representations from Transformers (BERT) [22]. After that, point placement-based visualization techniques are used to represent graphically each text document of the corpus in the visual space (layout). Finally, the quality of the obtained layouts are evaluated using metrics that measures the groups separation and neighborhood preservation.

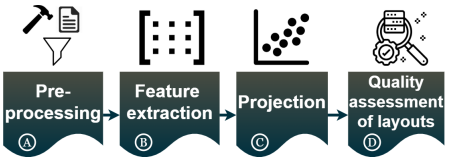

Fig. 1. Overview of the proposed approach.

A. Text preprocessing

Initially, a text preprocessing step is performed, which consists of: converting all uppercase characters to lowercase; tokenizing the text using NLTK's [23] recommended word tokenizer; retaining only words in which all characters are letters of the alphabet; and applying lemmatization and groups the tokens to form the text document. The removal of stopwords or more frequent words was performed internally by each technique.

B. Feature extraction

The feature extraction aims to generate feature vectors from the preprocessed text in order to enable the use of the visualization techniques and the classification models. The first text representation technique is tf-idf, which computes a feature vector for each document.

It is worth noting that word embedding techniques output a single vector of real values for each word of the corpus. These vectors present a fixed size and concentrate information regarding semantic relationships which allows to associate the context to its meaning. As text documents contain different numbers of words, we follow the strategy described by [24], [25] to obtain a single vector representing each text document. For that purpose, the "document vector" is obtained by averaging the word vectors and weighting them according to the frequency of each word. The same strategy is employed for the obtained global vector using GloVe.

In BERT, the idea is to consider transfer learning by means of pre-trained "bert-base-uncased" model provided by the Hugging Face transformers. The contextual embeddings of the last layer were extracted by disregarding the fine-tuning of any BERT parameters. This strategy is similar to that presented by Devlin et al. [26], which concluded that the performance obtained by concatenating the token representations of the top

four hidden layers of the pre-trained Transformer, using them directly in the downstream task, is comparable to that fine-tuning the entire model (including the BERT parameters). Prior to the model's training, the corpus was prepared by including two special tokens in the text: a token "[SEP]" to separate two sentences and a classification token "[CLS]" which refers to the first token of each tokenized sequence.

C. Visualization based on point placement

As a result of the previous step, the documents of the input corpus were transformed to feature vectors, which defines a high dimensional space. The goal of this step is to map the multidimensional instances to a bi-dimensional visual space (layout) so that we can visualize the similarity relations between text documents.

For that purpose, we consider four multidimensional projection techniques: Principal Component Analysis (PCA), Isometric Feature Mapping (Isomap), Uniform Manifold Approximation and Projection (UMAP) and Locally Linear Embedding (LLE). Additionally, the visualization technique t-distributed stochastic neighbor embedding (t-SNE) is also employed due to its previous successful applications in visual text analysis processes.

D. Quality assessment of layouts

The generated layouts are then evaluated using two state-of-art metrics that can evaluate the neighborhood preservation, cluster separation and the similarity preservation among instances. These metrics were selected since we are following the quality assessment strategy presented in related researches [16] [27] [28].

The trustworthiness is a metric based on neighborhood preservation which expresses for each instance in the original space, the proportion of k-nearest neighbor points that are retained in its k-neighbor points in the visual space. For each k, we compute this reliability by averaging the precision for all text instances. Values close to one indicate higher preservation of local structure of the original space in the layout.

The separation of grouped points in the layout is evaluated using the silhouette coefficient. For an instance x, the cohesion $a(x)$ is computed according to the mean intra-cluster distances, while the separation $b(x)$ is obtained by the minimum distance between x to any other instance belonging to another cluster. Eq. (1) presents the silhouette coefficient $s(x)$ for an instance x:

$$s(x) = \frac{b(x) - a(x)}{max(a(x), b(x))}, \tag{1}$$

in which we consider the average value of all silhouette coefficients regarding the instances in the corpus. Coefficient values closer to 1 indicate better cohesion and separation amongst clusters.

IV. Experimental results

In this section, we perform experiments in order to compare the quality of the layouts obtained by the visualizations based on point placement. The evaluation considers the metrics described in the previous section, as well as corpora of different aspects. The proposed method were coded in Python 3.8, in which the multidimensional projections were used from the scikit-learn [1] and umap [2] alongside plotly [3] to generate the graphics. The libraries gensim [4] and transformers were employed for feature extraction from text documents. The hyperparameters of the visualization techniques were adjusted based on visual analysis of the generated layouts and by considering the values of the layout evaluation metrics in some runs. Moreover, the classifier hyperparameters were set according to the results provided by the optimization approach *RandomizedSearchCV* [5].

In our experiments, we considered two public corpora from the Hugging Face library [29]: "amazon_polarity" [30] and "ag_news" [31]. The first one is appropriate to text and sentiment classification, and consists of amazon analytics over an 18-year period, including about 35 million analytics as of March 2013, product information and users, ratings etc. "ag_news" is a simple text review corpus for text classification tasks and it is defined by a collection of approximately 1 million news articles gathered from over 2000 news sources by "ComeToMyHead" in over a year of activity. As these corpora are very large, we subsample 7000 and 7600 documents from the "amazon_polarity" and the "ag_news", respectively.

A. Quality assessment of layouts

In order to study possible relations between the quality of layouts and the classification performance, we perform a classification evaluation using the corpora. In this sense, the low dimensional space generated for the text visualization is used as input to a classifier based on Support Vector Machines. For the sake of simplicity, we apply Holdout Cross Validation, in which $2/3$ of the data instances are used for training and the remaining are used for test.

Tables I, II, III and IV present the results of the silhouette coefficient (SC) and the F1-Score from the SVM classification on the test sets of both corpora regarding tf-idf, word2vec, BERT and GloVe combined with each visualization technique, respectively. The results related to "ag_news dataset" show a relationship between SC and classification accuracy, indicating that well-formed clusters are associated to higher F1-Scores. On the other hand, the results obtained using the "amazon_polarity" were affected by the overlapping of clusters in the low dimensional spaces.

Figures 2 and 3 show the preservation of the neighborhood by varying the number of nearest neighbors in relation to each data instance. In "amazon_polarity" corpus, word2vec presented the best precision scores and t-SNE also yielded satisfactory neighborhood preservation technique. However, in

[1] https://scikit-learn.org/stable/

[2] https://umap-learn.readthedocs.io/en/latest/

[3] https://plotly.com/

[4] https://radimrehurek.com/gensim/

[5] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

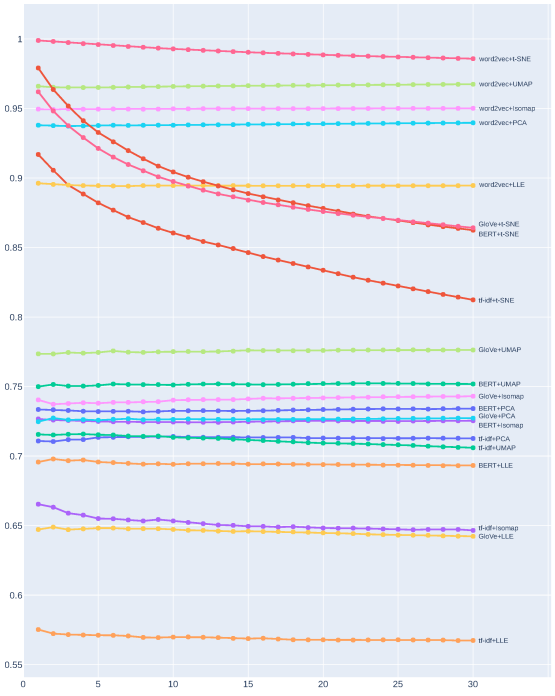

Fig. 2. Trustworthiness computed from the layouts generated by the visualization techniques using the "amazon_polarity" corpus.

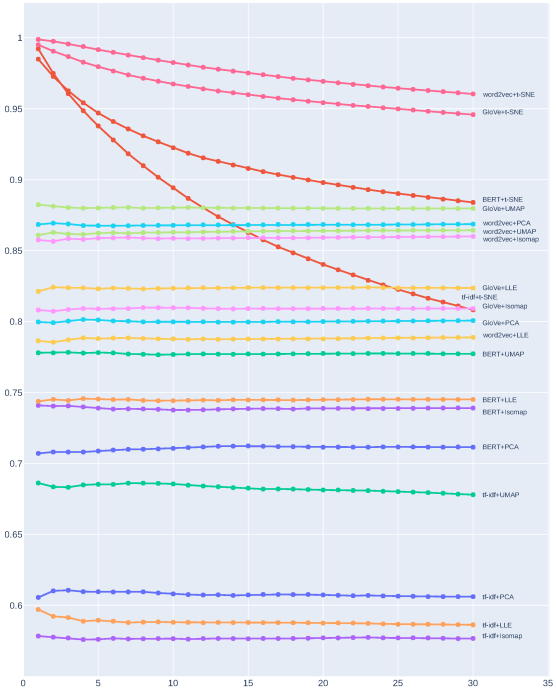

Fig. 3. Trustworthiness computed from the layouts generated by the visualization techniques using the "ag_news" corpus.

the "ag_news" corpus, word2vec and GloVe showed superior performances along with t-SNE.

In the layouts obtained by t-SNE, the decay of scores for increasing number of neighbors was expected since the algorithm aims to preserve only the closest neighbors of each point, thus effectively ignoring the most distant ones. Furthermore, as t-SNE is able to reflect the similarities between high dimensional data points, the satisfactory results regarding neighborhood preservation were expected.

It is important to note that word incorporation techniques word2vec and GloVe considerably affect the placement of dots in the layout and therefore have a strong influence on the layout generation. This can be verified in obtaining better reliability scores (neighborhood preservation) obtained when text documents are represented by these techniques, even when they are projected by LLE and Isomap, considered less efficient than other algorithms. Word2vec gave scores around 0.78 and 0.85 and GloVe gave scores around 0.82 and 0.80 when "ag_news" corpus are projected with LLE and Isomap respectively.

Techniques	amazon_polarity		ag_news	
	SC	f1-score	SC	f1-score
tf-idf + PCA	0.02	0.57	-0.12	0.23
tf-idf + t-SNE	0.00	0.47	0.21	0.80
tf-idf + UMAP	0.00	0.51	0.21	0.79
tf-idf + Isomap	0.00	0.34	0.19	0.55
tf-idf + LLE	0.00	0.34	-0.03	0.09

TABLE I
SILHOUETTE COEFFICIENT (SC) AND F1 (CLASSIFICATION) SCORES FOR TF-IDF.

Techniques	amazon_polarity		ag_news	
	SC	f1-score	SC	f1-score
word2vec + PCA	0.01	0.34	-0.04	0.24
word2vec + t-SNE	0.01	0.52	-0.03	0.34
word2vec + UMAP	0.01	0.53	-0.03	0.28
word2vec + Isomap	0.01	0.53	-0.04	0.27
word2vec + LLE	0.01	0.53	-0.05	0.09

TABLE II
SILHOUETTE COEFFICIENT (SC) AND F1 (CLASSIFICATION) SCORES FOR WORD2VEC.

Techniques	amazon_polarity		ag_news	
	SC	f1-score	SC	f1-score
BERT + PCA	0.00	0.34	0.04	0.46
BERT + t-SNE	0.01	0.60	0.21	0.79
BERT + UMAP	0.01	0.56	0.20	0.77
BERT + Isomap	0.01	0.55	0.18	0.72
BERT + LLE	0.00	0.34	0.18	0.09

TABLE III
SILHOUETTE COEFFICIENT (SC) AND F1 (CLASSIFICATION) SCORES FOR BERT.

Techniques	amazon_polarity		ag_news	
	SC	f1-score	SC	f1-score
GloVe + PCA	0.00	0.34	0.22	0.69
GloVe + t-SNE	0.01	0.55	0.27	0.84
GloVe + UMAP	0.01	0.54	0.26	0.82
GloVe + Isomap	0.00	0.34	0.26	0.76
GloVe + LLE	0.01	0.34	0.23	0.09

TABLE IV
SILHOUETTE COEFFICIENT (SC) AND F1 (CLASSIFICATION) SCORES FOR THE GLOVE.

B. Visual analysis of text collections

Figure 4 depicts the layouts obtained by the visualization techniques PCA, t-SNE, UMAP, Isomap and LLE using 7000

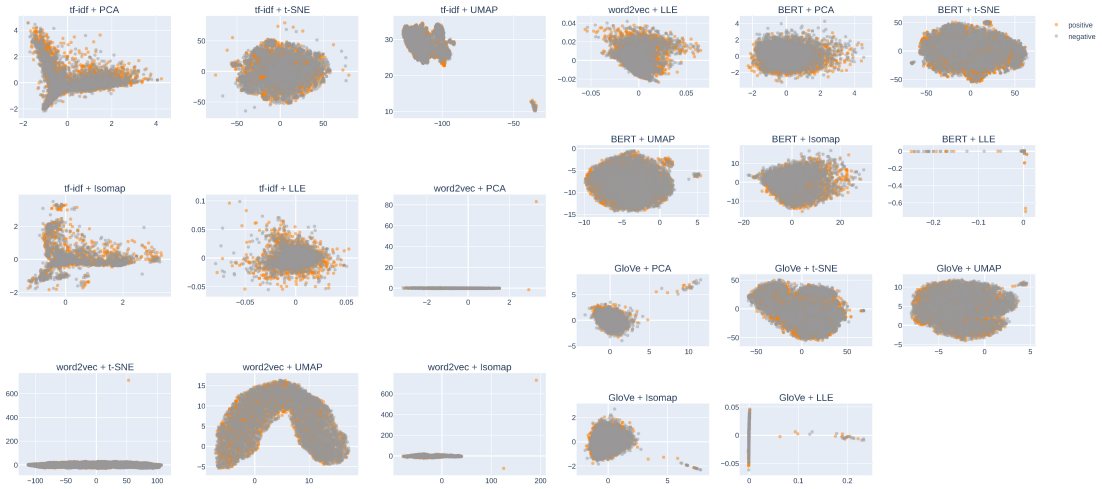

Fig. 4. Layouts produced by the visualizations PCA, Isomap, LLE, t-SNE and UMAP using the "amazon_polarity" in different representations.

Fig. 5. Layouts produced by the visualizations PCA, Isomap, LLE, t-SNE and UMAP using the "ag_news dataset" in different representations.

instances of the corpus "amazon_polarity" represented by tf-idf, word2vec, GloVe and BERT. It is possible to verify two groups of positive and non-positive comments of the corpus presenting overlap in the layout. Therefore, it is not possible to identify the best layout in advance by visual analysis.

In Figure 5 with the projections of 7600 instances of the corpus "ag_news", the visual analysis allows to identify the separation of classes visually better in the projections of word embedding tf-idf, BERT, and GloVe. The results obtained using word2vec can be improved to by incorporating a Neural Network layer on top of word vectors of a document to combine them. This method can also be extended to GloVe.

We can also conclude that f1-score is more strongly related to SC than to neighborhood preservation. For instance, we verify in the "amazon_polarity" corpus that the best sorting accuracies, with f1-score above 0.5, also had the best neighborhood preservation. However, in the "ag_news dataset", the combination of word embedding techniques, BERT and GloVe, with the most successful visualization techniques according to the literature, t-SNE and UMAP, yielded in better defined clusters in the associated layouts and, consequently, higher f1 scores in the evaluation of classification performance.

The low performance of PCA in the classification tasks and layout qualities for both corpora can be explained by the common local structures of feature spaces obtained by the document representations, thus affecting the capture of data variability in two principal components.

V. Conclusion

This paper described a study to compare visualization to explore feature spaces of different text document representation. Visual representations can be considered as a guide for understanding the behavior of features in terms of the similarity or dissimilarity of textual documents. Various word embedding techniques along with point placement-based visualizations were analyzed and compared in relation to the quality of the resulting layouts and a classification task.

The experimental results showed that layouts depicting grouped points, especially those presenting higher silhouette scores, are associated to superior rankings. Furthermore, it was possible to represent document by feature vectors obtained from word embeddings, since BERT, word2vec and GloVe achieved satisfactory layout quality when employed with powerful visualizations, such as t-SNE.

Future work can be guided to incorporate other techniques and word embedding models such as Doc2Vec [32], GPT [33], RoBERTa [34], ELMo [35] etc. Additional metrics for layout quality assessment will be investigated, such as Shapley values and Neighborhood Hit. Finally, this research will explore an interactive visual exploration tool of text collections, retaining user control and allowing users to transform the feature space

ACKNOWLEDGMENT

We would like to thank Conselho Nacional de Pesquisa e Desenvolvimento (CNPq) (process #115371/2021-4) for providing the undergraduate grant that supported this research.

REFERENCES

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[2] L. G. Nonato and M. Aupetit, "Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 8, pp. 2650–2673, 2018.

[3] N. Saeed, H. Nam, M. I. U. Haq, and D. B. Muhammad Saqib, "A survey on multidimensional scaling," *ACM Computing Surveys (CSUR)*, vol. 51, no. 3, pp. 1–25, 2018.

[4] S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Information Fusion*, vol. 59, pp. 44–58, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S156625351930377X

[5] M. Espadoto, R. M. Martins, A. Kerren, N. S. T. Hirata, and A. C. Telea, "Toward a quantitative survey of dimension reduction techniques," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 3, pp. 2153–2173, 2021.

[6] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural computation*, vol. 11, no. 2, pp. 443–482, 1999.

[7] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

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

[9] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[10] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[11] L. Huang, S. Matwin, E. J. de Carvalho, and R. Minghim, "Active learning with visualization for text data," in *Proceedings of the 2017 ACM Workshop on Exploratory Search and Interactive Data Analytics*, 2017, pp. 69–74.

[12] L. Hajderanj, I. Weheliye, and D. Chen, "A new supervised t-sne with dissimilarity measure for effective data visualization and classification," in *Proceedings of the 2019 8th International Conference on Software and Information Engineering*, 2019, pp. 232–236.

[13] Y. Han, Z. Wang, S. Chen, G. Li, X. Zhang, and X. Yuan, "Interactive assigning of conference sessions with visualization and topic modeling," in *2020 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 2020, pp. 236–240.

[14] K. Kim and J. Lee, "Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction," *Pattern Recognition*, vol. 47, no. 2, pp. 758–768, 2014.

[15] Y. Zhao, B. Qin, T. Liu, and D. Tang, "Social sentiment sensor: a visualization system for topic detection and topic sentiment analysis on microblog," *Multimedia Tools and Applications*, vol. 75, no. 15, pp. 8843–8860, 2016.

[16] R. Motta, R. Minghim, A. de Andrade Lopes, and M. C. F. Oliveira, "Graph-based measures to assist user assessment of multidimensional projections," *Neurocomputing*, vol. 150, pp. 583–598, 2015.

[17] D. Smilkov, N. Thorat, C. Nicholson, E. Reif, F. B. Viégas, and M. Wattenberg, "Embedding projector: Interactive visualization and interpretation of embeddings," *arXiv preprint arXiv:1611.05469*, 2016.

[18] S. Liu, P.-T. Bremer, J. J. Thiagarajan, V. Srikumar, B. Wang, Y. Livnat, and V. Pascucci, "Visual exploration of semantic relationships in neural word embeddings," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 553–562, 2017.

[19] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008, vol. 39.

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[21] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[22] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[23] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[24] A. Rexha, M. Kröll, M. Dragoni, and R. Kern, "Polarity classification for target phrases in tweets: a word2vec approach," in *European Semantic Web Conference*. Springer, 2016, pp. 217–223.

[25] O. Abdelwahab and A. Elmaghraby, "Uofl at semeval-2016 task 4: Multi domain word2vec for twitter sentiment classification," in *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 2016, pp. 164–170.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[27] J. G. S. Paiva, W. R. Schwartz, H. Pedrini, and R. Minghim, "Semi-supervised dimensionality reduction based on partial least squares for visual analysis of high dimensional data," in *Computer Graphics Forum*, vol. 31, no. 3pt4. Wiley Online Library, 2012, pp. 1345–1354.

[28] A. Chatzimparmpas, R. M. Martins, and A. Kerren, "t-visne: Interactive assessment and interpretation of t-sne projections," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 8, pp. 2696–2714, 2020.

[29] "Hugging face - datasets," https://huggingface.co/docs/datasets.

[30] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," 2016.

[31] A. Gulli, "Ag's corpus of news articles," http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html.

[32] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*. PMLR, 2014, pp. 1188–1196.

[33] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[34] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[35] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.