Document Clustering, Summarization, and Visualization

Prateek Pandey
Arizona State University
Tempe, USA
prateek.pandey@asu.edu

Jayavardhan Karampudi
Arizona State University
Tempe, USA
jkarampu@asu.edu

Akashkiran Shivakumar
Arizona State University
Tempe, USA
ashivak4@asu.edu

Raviram Mamidi
Arizona State University
Tempe, USA
rmamidi2@asu.edu

Sundaravadivel CP
Arizona State University
Tempe, USA
sundarav@asu.edu

Tejesh Andhavarapu
Arizona State University
Tempe, USA
tandhava@asu.edu

***Abstract*—This project aims to investigate and apply various clustering and visualization approaches to textual documents. Modern clustering methods will be applied to new data sets, and the results will be presented using Uniform Manifold Approximation and Projection (UMAP). The Universal Sentence Encoder will generate sentence embed-dings for the text. These documents will be clustered using techniques such as K-Means, HDBSCAN, and LDA (Latent Dirichlet Allocation) on the produced embedding vectors. The proposed solution groups comparable documents based on the generated embedding and gives a graphical representation of these articles. Finally, Sentiment Analysis is performed using the BART Facebook encoder-decoder model and Spacy, with the results shown.**

***Index Terms*—Document clustering, Summarization, HDB-Scan, LDA, Agglomerative clustering, Visualization.**

I. Introduction

Clustering touches humans in all aspects of life, from the brain's neuronal activity to the way it perceives patterns to physically grouping physical facts for simplicity of calculation and duplication. Clustering has been the subject of continuing research in a variety of fields, including statistics, pattern recognition, and machine learning. Clustering is a data mining technique used to handle very large datasets with varying data attributes. As a result, the performance of the clustering approaches is constrained in numerous ways. Many new algorithms have recently been developed and successfully applied to real-world data mining problems.

Deep learning advances in recent years have greatly improved algorithms' ability to analyze text. Modern artificial intelligence algorithms used wisely can be a helpful tool for deciphering a person's emotions from textual data.

II. Problem Statement

1) With the development of the internet, a large number of documents are now accessible online, making it challenging to locate and extract crucial information.
2) Summarizing an extensive amount of text is challenging and time-consuming. Calculations and extensive text processing are needed.
3) A set of documents are grouped using document clustering, which uses a similarity score. When clustering is integrated with a search engine, we can view the document set's overall structure and delve as deeply as we like into it.
4) Document summarization helps in gaining a subjective understanding of the articles while saving a lot of time.
5) The main goal of the project is to
    a) Group the articles together and give a succinct summary
    b) Utilize visualization strategies to demonstrate the relevance
    c) Document Summarization

III. Related Works

In the past, the two most commonly used algorithms for clustering tasks were K-means and DBScan. Initially, these algorithms were considered for clustering documents. However, after conducting further research, it became evident that K-means, which strictly assigns one label or category to a group, would not be suitable for document clustering. This is because a document can belong to multiple categories. To address this, we explored Latent Dirichlet Allocation (LDA) [1], which considers this aspect and provides a probabilistic composition of the document, producing a probability distribution of groupings (topics) per document. Another algorithm we used was HDBSCAN [13], an improvement over DBSCAN, which employs a hierarchical clustering algorithm suitable for multi-dimensional data representation. A hybrid strategy was employed in the Scatter/Gather system [18], a clustering-based document browsing system for the textual domain, which included K-means and Agglomerative hierarchical clustering.

Identify applicable funding agency here. If none, delete this.

Later on, we will elaborate on the reason for our dataset choice, but for our text classification tasks, we utilized the 20 newsgroup dataset. The bag of words model is a popular and preferred feature model in literature due to its simplicity and high performance in various text classification tasks. This model represents the text to be classified as a collection of individual words with no correlation between them, completely disregarding grammar and word order within the text. It has been widely used in sentiment analysis and has shown good performance despite its simplifying assumptions. One way to utilize the prior polarity of words as features is through the use of publicly available lexicons or dictionaries that map words to their prior polarity. This unsupervised approach is relatively simple to implement.

IV. System Architecture And Algorithms

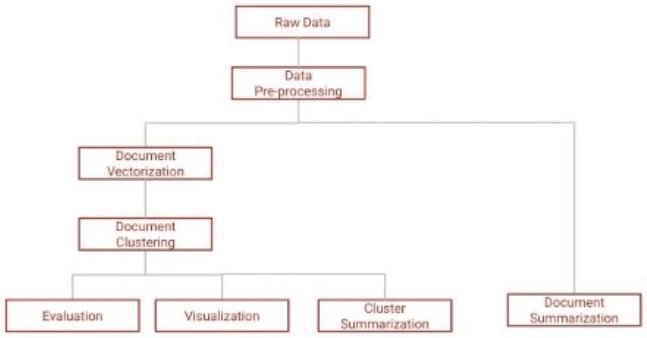

Fig. 1. System Architecture

A. Document Clustering

1) Latent Dirichlet Allocation: Long documents can be automatically categorized and understood using the Latent Dirichlet Allocation[1] method. Each topic has a different set of terms, and documents are made up of a variety of words. LDA analyzes a document's words to determine the topics to which each word belongs.In some ways, LDA and k-means clustering are similar: both are unsupervised learning techniques that don't need pre-trained data. LDA, on the other hand, assigns multiple topic labels, allowing for a more nuanced understanding of the data, whereas k-means clustering only assigns a single topic label to each data set.

2) HDBScan: The hierarchical clustering algorithm HDBSCAN [13] is capable of capturing both the shape or property of clusters and their density while also taking into account noise.The DBSCAN algorithm is implemented by HDBSCAN, which does not require a predetermined distance threshold because it accepts a range of epsilon values. By transforming DBSCAN into a hierarchical clustering algorithm and then employing a method to extract a flat clustering based on the stability of clusters, HDBSCAN extends DBSCAN.

3) Agglomerative Clustering: Each data point is initially treated as a separate cluster in the agglomerative hierarchical clustering [14] technique. After that, clusters are combined to create a new cluster. This procedure is repeated until a single cluster is formed from all of the data points. As a bottom-up method, agglomerative hierarchical clustering creates larger clusters with each iteration of the algorithm. You don't need to specify how many clusters you want to divide the data into, in contrast to k-means clustering. To determine how many clusters should be present in a dataset, use agglomerative hierarchical clustering.

B. Summarization

There are two ways to summarize a text. The first method, known as extractive summarization, aims to pinpoint the key phrases and use them directly as the summary. In this approach, stop words and punctuation are first removed after tokenizing the words. The frequency and weights of each word in the text are then determined. The weight assigned to each word is the ratio of each word's frequency to the maximum of frequencies of all the words within the document.The total weight for each sentence is determined after the weights have been assigned. The sentences with higher weights are regarded as being more significant and are added to the end of the summary. However, this method did not produce a very meaningful summary, so we switched to an advanced method called abstractive summarization. In abstractive summarization, the document's context or meaning are taken into account when they create a new summary. The summary may contain entirely original sentences that accurately and meaningfully summarize the substance of the paper. This strategy has been put through practice using Facebook BART Large CNN [19]. The BART model can be utilized for text comprehension problems and is especially effective when tuned for text generation. In the experiment, we used a BART model that had already been trained and was enhanced using CNN Daily Mail, a sizable database of text-summary pairings. We loaded the pre-trained model and applied its summarizer to the loaded documents to provide an insightful summary. Additionally, a large document was created by appending the text from every document in a cluster and summarizing it. The resulting summary is an overview of the whole cluster.

V. Dataset

A. Dataset Description

In this project, we examined several publicly available datasets on Kaggle [5]. Initially, we considered the Reuters dataset, but we found that it had multiple categories that overlapped and were non-exhaustive. Additionally, there were relationships among the categories, making it more suitable for a classification problem rather than a clustering problem. Subsequently, we explored a medium articles dataset, but it contained a significantly lower number of documents than we required for clustering. Eventually, we settled on the 20 newsgroups dataset [2], which is a collection of 20,000 documents evenly distributed across 20 different newsgroups, each pertaining to a distinct topic. Unlike the Reuters dataset, the categories in this dataset are not highly related, and it contains a substantial number of documents, making it the most appropriate fit for our problem statement.

B. Data Pre-processing and Vectorization

As a part of our data pre-processing, we first analyzed the 20 newsgroup dataset and extracted all the subsets of fetch 20 newsgroups. We removed the headers and footers from the dataset and extracted the text and labels into data frames. Subsequently, we tokenized the sentences into individual tokens and converted them into lowercase alphabets, removing punctuation, white spaces, and stop words. Lastly, we applied lemmatization to normalize the sentence and preserve its meaning. We then converted the data into vector form, removing all null rows in the process [7]. The resulting pre-processed data consisted of 18864 rows and 2 dimensions. Next, we passed this pre-processed data to Google's universal sentence encoder to generate sentence embeddings, which were later utilized by the clustering algorithms. These embeddings allowed us to cluster the meanings of entire sentences, thereby reducing the amount of training data required to achieve effective clustering results, as compared to clustering individual words.

VI. EVALUATIONS

Evaluation of the clustering algorithm performance is crucial after clustering is complete. In order to do this, we have selected five measures that fit this problem statement the best.

A. Homogeneity

This metric gauges how similar a sample's members are to one another. It is between 0 and 1. For instance, the homogeneity would be 1 if all samples that are a part of cluster k were given the same label, "c."

B. Completeness

This metric assesses how frequently the clustering algorithm groups samples that share characteristics. It is between 0 and 1. For instance, completeness would be 1 if all samples with label "c" were placed in the same cluster.

C. Adjusted Rand Index

This metric establishes if two clusters are comparable to one another. It is between 0 and 1. Random labeling is indicated by a 0 adjusted rand index, and identical partitioning is indicated by a 1.

D. V-measure

The V-measure evaluates how well a partition for clustering works. It is the homogeneity and completeness harmonic average. The V- measure will likewise be low if homogeneity or completeness are both low.

E. Silhouette Score

This metric is used to determine a metric's quality clustering technique. It ranges from -1 to 1. Silhouette score is 1 when the clusters are well apart and clearly distinguished. This score is 0 when the clusters are indifferent and -1 when the clusters are assigned in a random fashion.

All of the clustering methods listed in the algorithms section have had the aforementioned five assessment metrics calculated for them, as shown in Table 1. In comparison to HDBScan and agglomerative clustering, LDA outperformed them in four of the five metrics tested: homogeneity, completeness, v-measure, and adjusted rand index. In contrast, HDBScan outperformed it in terms of the evaluation metric known as the Silhouette score. Finally, we deduce that the LDA clustering algorithm was the most effective one for this dataset.

VII. VISUALIZATIONS

We have visualized the clusters using t-Distributed Stochastic Neighbor Embedding (t-SNE) [17], Uniform Manifold Approximation and Projection (UMAP) [15], Compression Variational Autoencoder(CVAE) [12].

A. t-SNE

The dimensionality reduction technique t-Distributed Stochastic Neighbor Embedding (t-SNE) is especially well adapted to the visualization of high-dimensional datasets [16]. Through t-SNE, the data is displayed in a smaller dimension while preserving its local organization. A cost function is utilized to calculate a similarity measure between pairs of instances in both the high-dimensional and low-dimensional space in order to maximize the similarity measures.

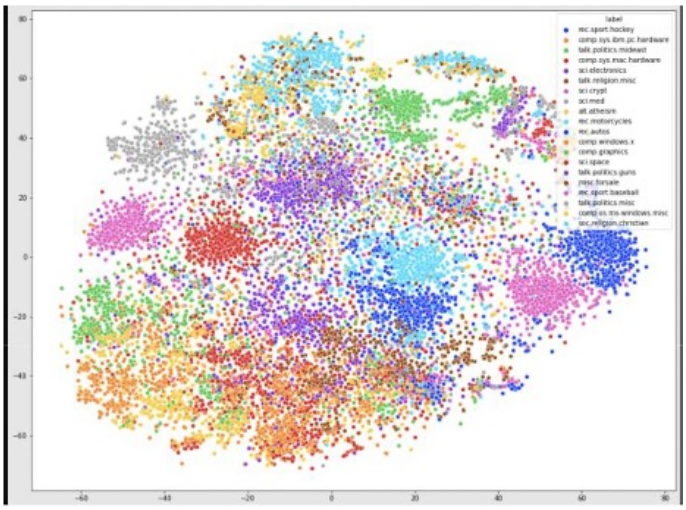

Fig. 2. Visualization of LDA clusters using t-SNE

B. UMAP

Similar to t-SNE, a dimension reduction technique known as Uniform Manifold Approximation and Projection (UMAP) [8] can be used for nonlinear general dimension reduction and visualization. UMAP is quicker than t-SNE since it makes use of many optimization techniques to quicken the process. UMAP can be used to the dataset without dimensionality reduction and data pre-processing.

C. CVAE

Compression In addition to building on the simplicity of the t-SNE and UMAP implementations, the Variational Autoencoder (VAE) [9] introduces a number of highly desirable characteristics. Since it is based on variational autoencoders, it is faster than t-SNE or UMAP. Contrary to t-SNE and UMAP, which perform better even with smaller sets of data, CVAE scales well to high dimensional input and latent spaces despite requiring a large amount of data to be trained. Furthermore, CVAE frequently only achieves a weak separation between clusters.

Clustering Technique	Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Silhouette Coefficient
LDA	0.583	0.584	0.584	0.491	0.014
HDBScan	0.317	0.493	0.385	0.132	0.343
Agglomerative Clustering	0.379	0.396	0.387	0.206	0.004

Fig. 3. Evaluation Results of Clustering Algorithms

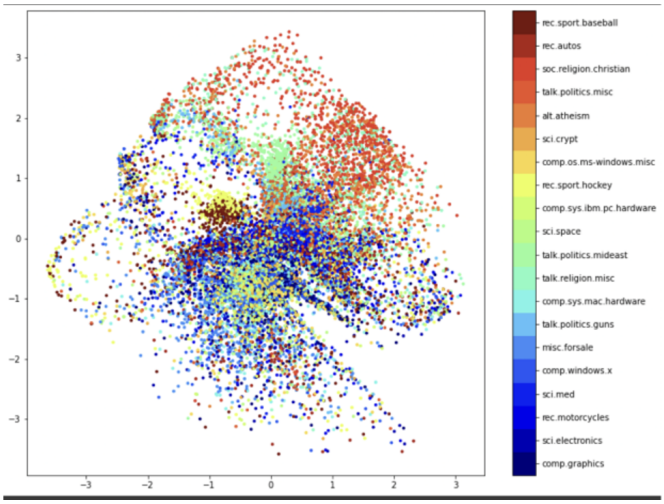

Fig. 4. Visualization of HDBScan clusters using t-SNE

VIII. TEAM CONTRIBUTION

The contribution of the team members can be outlined below

A. Study of clustering and visualization techniques

Conduct research on various clustering and visualization techniques for use on datasets. The algorithms to be utilized have been researched by all team members.

B. Data Pre-processing

Sundar, Tejesh, and Prateek are in charge of data preprocessing. The data was preprocessed by reducing noise and transferring it to a process for data embedding.

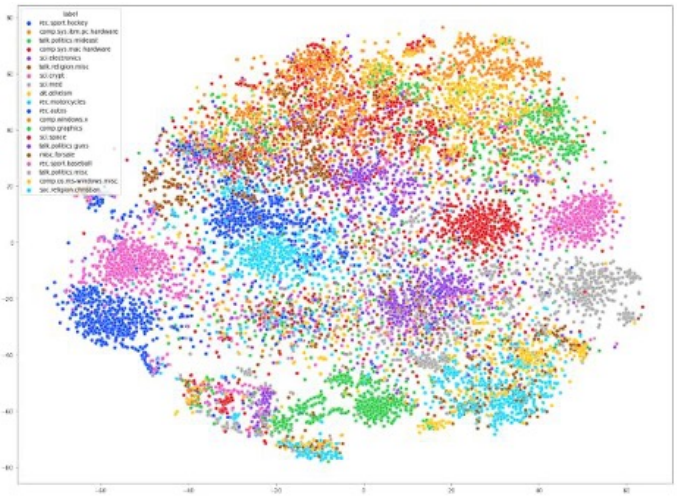

Fig. 5. Visualization of Agglomerative clusters using t-SNE

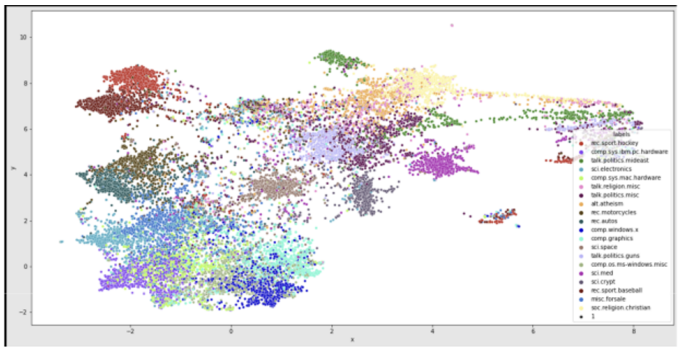

Fig. 6. Visualization of LDA clusters using UMAP

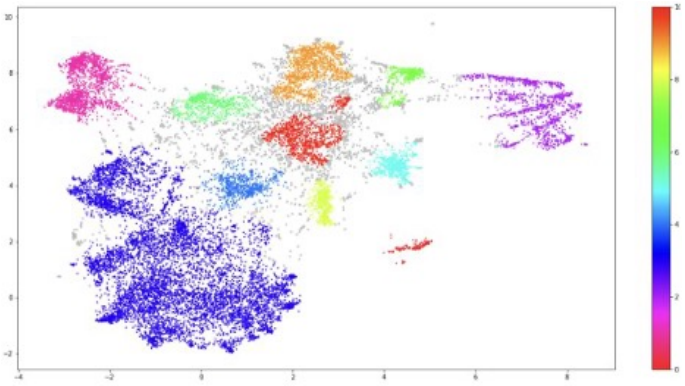

Fig. 7. Visualization of HDBScan clusters using UMAP

Task	Team Members	Deadline
Study of clustering and visualization techniques	All team members	Jan 31 - Feb 16
Data Pre-processing	Sundar,Tejesh, Prateek	Feb 17 - Feb 28
Data Embedding	Akash Kiran, Jayavardhan, Raviram	Feb 28 - March 15
Clustering	All team members	Mar 15 - Mar 30
Document summarization	All team members	April 1 - April 5
Visualization	Prateek,Raviram, Jayavardhan	April 5 - April 10
Summary, Final Evaluation and Analysis	All team members	April 10– April 15

Fig. 8. Contributions of Each Team Members

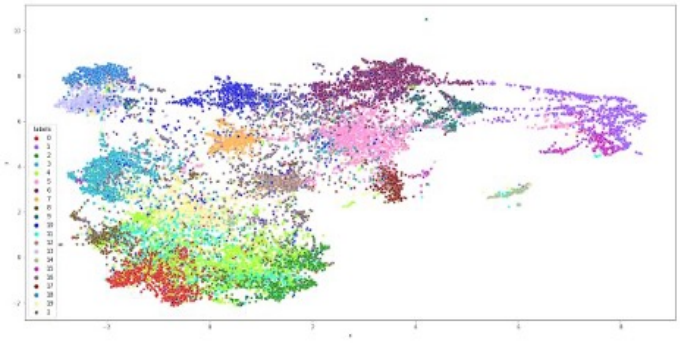

Fig. 9. Visualization of Agglomerative clusters using UMAP

C. Data Embedding

Akash Kiran, Jayavardhan, and Raviram have been in charge of data embedding. To generate sentence encoding, the preprocessed data were transformed to vector form using Google's universal sentence encoder.

D. Clustering

To cluster the papers, we investigated various clustering algorithms. We then completed LDA, agglomerative clustering, and HDBScan. This task involved the participation of all team members

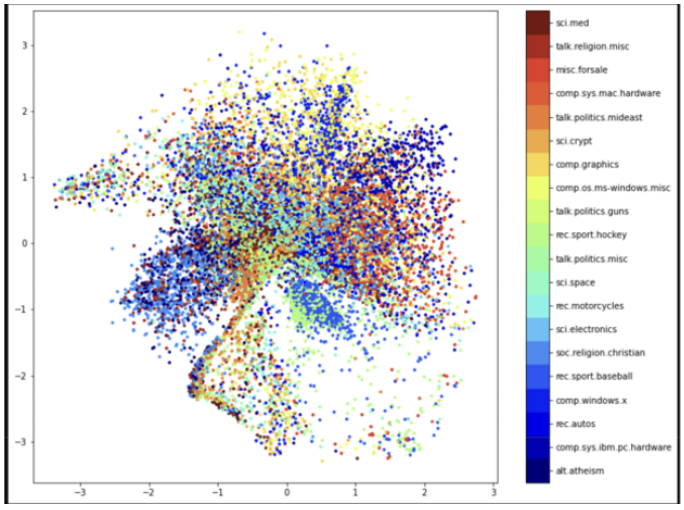

Fig. 10. Visualization of LDA clusters using CVAE

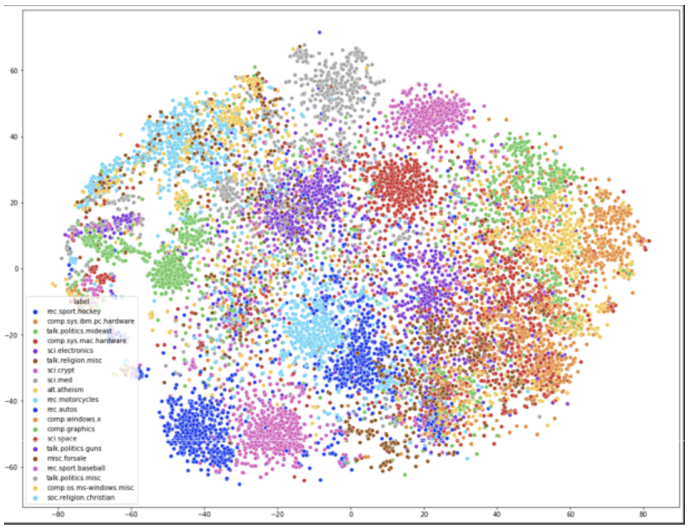

Fig. 11. Visualization of HDBScan clusters using CVAE

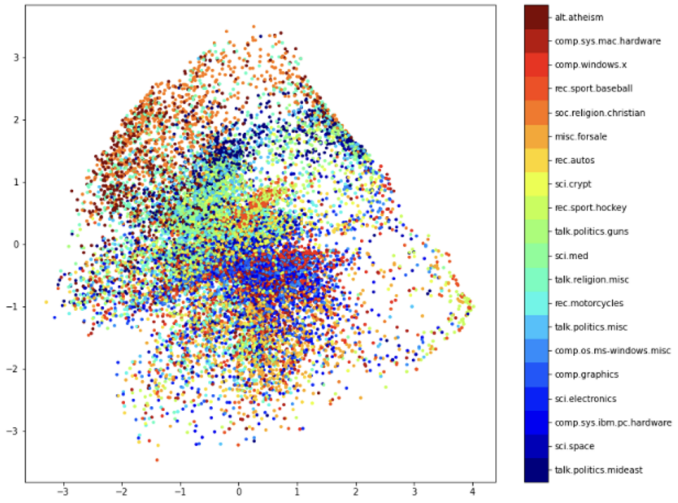

Fig. 12. Visualization of Agglomerative clusters using CVAE

E. Document Summarization

We used Abstractive and Extractive text summary techniques to construct two methods of summarization. All of the team members have contributed to the summarization component.

F. Document Visualization

Prateek, Raviram, and Jayavardhan were the ones in charge of the module that implemented the document visualization utilizing UMAP, T-SNE, and Compression VAE.

G. Evaluation and Analysis and Summary and Documentation

All team members were responsible for determining appropriate evaluation metrics for document clustering and analyzing the results.. We have completed the evaluation metrics of

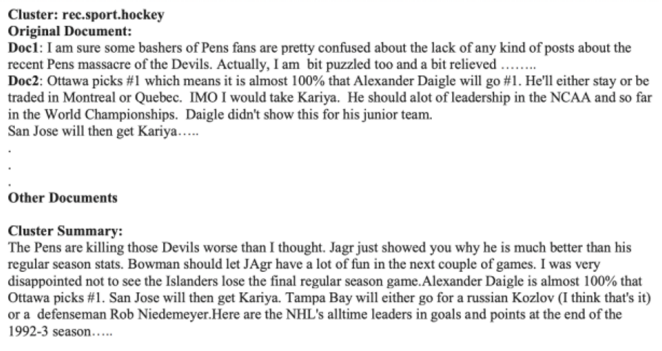

Fig. 13. Output of document summarization of a cluster

homogeneity, completeness, V-measure, adjusted Rand-Index, and silhouette coefficient. All team members contributed to the project report by documenting various sections of it.

IX. CONCLUSION

In conclusion, we carried out a subjective analysis of the dataset throughout this research by visualizing the clusters, which allowed us to do so. We were able to do clustering using methods like LDA, HDBScan, and K-Means. The grouped categories were then visualized using UMAP and t-SNE. A comparison of the algorithms was done using the evaluation measures that we developed. It was evident that LDA was delivering the most accurate outcomes conceivable given the dataset at hand. Using SpaCy [20], we performed extractive text summarization [10]. We loaded the model and used word frequency to determine weights. Then, we take that vector and use Facebook's BART large CNN to create a summary using a sentence token. next using the cluster, we summarize the sentence in the data frame in fewer than 2700 characters. An encoder-decoder transformer is what the Facebook BART is. having features like BERT and GPT3 that are comparable to or better than the T5 model. As a result, the summary is state-of-the-art, and we verified that it provided an accurate overview of the subjects. Along with being able to discern the relationships between related categories in the dataset.

REFERENCES

[1] http://qwone.com/ jason/20Newsgroups/.
[2] https://albertauyeung.github.io/2020/06/19/bert-tokenization.html/.
[3] Millar, Jeremy R. et al. "Document Clustering and Visualization with Latent Dirichlet Allocation and Self-Organizing Maps." FLAIRS Conference (2009).
[4] arXiv:2012.04456 "Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap,and PaCMAP for Data Visualization".
[5] Ceran, B., Kedia, N., Corman, S.R., Davulcu, H., 2015, Story Detection Using Generalized Concepts and Relations, Proceedings of International Symposium on Foundation of Open Source Intelligence and Security Informatics (FOSINT-SI), in conj. with IEEE ASONAM 2015, Paris, France
[6] https://towardsdatascience.com/compressionvae-a-powerful-and-versatile-alternative-to-t-sne-and-umap-5c50898b8696.
[7] Agglomerative-hierarchical-clustering - https://medium.com/geekculture/agglomerative-hierarchical-clustering-a-gentle-intro-with-an-example-program-4b7afe35fd4b.
[8] https://lvdmaaten.github.io/tsne/.
[9] Lloyd,S.P.(1957).LeastsquaresquantizationinPCM.TechnicalReport RR-5497, Bell Lab, September 1957.
[10] https://aparnamishra144.medium.com/automated-text-summarization-using-spacy-in-nlp-8750b1b6e404
[11] Giri. (2021, May 2). Is Latent Dirichlet Allocation (LDA) A clustering algorithm? HDS; High Demand Skills. https://highdemandskills.com/lda-clustering/.
[12] https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection.
[13] Karmakar, Saurav. "Syntactic and Semantic Analysis and Visualization of Unstructured English Texts." (2011).
[14] Cao,Tuan-Dungetal."HotTopicDetectiononNewspaper"Conference: the Ninth International Symposium (2018)
[15] Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap,and PaCMAP for Data Visualization
[16] van der Maaten, L.J.P. ; Hinton, G.E. / Visualizing High-Dimensional Data Using t-SNE. In: Journal of Machine Learning Research. 2008 ; Vol. 9, No. nov. pp. 2579-2605.
[17] HDBScan - https://towardsdatascience.com/tuning-with-hdbscan-149865ac2970.
[18] https://umap-learn.readthedocs.io/en/latest/.
[19] https://towardsdatascience.com/compressionvae-a-powerful-and-versatile-alternative-to-t-sne-and-umap-5c50898b8696.
[20] BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension