Document Clustering, Summarization, and Visualization

Akash Kant
Arizona State University
Tempe, USA
akant1@asu.edu

Mounika Chadalavada
Arizona State University
Tempe, USA
mchadala@asu.edu

Yasir Affan
Arizona State University
Tempe, USA
yaffan@asu.edu

Sagar Sharma
Arizona State University
Tempe, USA
sshar200@asu.edu

Swaapnika Motapothula
Arizona State University
Tempe, USA
smotapot@asu.edu

Harshitha Sadula
Arizona State University
Tempe, USA
hsadula@asu.edu

***Abstract*—This project aims to explore and implement various clustering, and visualization techniques on textual documents. State-of-the-art algorithms to cluster documents will be applied to new data sets and results will be visualized using Uniform Manifold Approximation and Projection (UMAP). Sentence Embeddings will be generated for the text using the Universal Sentence Encoder. For clustering, these documents, techniques like K-Means, HDBSCAN, and LDA (Latent Dirichlet Allocation) will be used on the generated embedding vectors. The proposed solution clusters similar documents based on the embedding generated and provides a graphical visualization for these articles. At last, Sentiment Analysis is also done using BART Facebook encoder-decoder model and Spacy and the results are visualized. Index Terms—Latent Dirichlet Allocation (LDA), Uniform Manifold Approximation and Projection (UMAP), HDBSCAN, Document Clustering, Universal Sentence Encoder, Document Visualization.**

***Index Terms*—document clustering, HDBScan, LDA, agglomerative clustering, visualization, summarization**

I. Introduction

Humans deal with clustering in all facets of life, from the brain's neuronal activity to the way it recognizes patterns to actually grouping physical data for ease of calculation and replication. It should come as no surprise that clustering has been the focus of ongoing study in a number of domains, including statistics, pattern recognition, and machine learning. Clustering is a technique used in data mining to handle very big datasets with various data properties. As a result, the performance of the clustering methods is subject to several constraints. Many new algorithms have lately been developed and effectively used to solve actual data mining issues. Recent developments in deep learning have significantly increased algorithms' capacity for text analysis. Intelligent use of cutting-edge artificial intelligence algorithms can be a useful tool for analyzing a person's feelings from textual data.

II. Problem Statement

1) With the evolution of the internet, many documents are available online and it has been difficult to find out and extract important information.
2) Large-scale text summarization is difficult and time-consuming. Extensive text processing and calculations are required.
3) Document clustering is grouping a set of documents based on a similarity score. Integrated with any search engine, clustering allows us to see the overall structure of the document set and browse as deep into it as you want.
4) Document summarization saves a lot of time and helps in gaining a subjective understanding of the articles.
5) The main goal of the project is to
   - Cluster the articles and provide a short summary
   - Apply visualization techniques to showcase relevancy
   - Document Summarization

III. Related Works

The most popular algorithms used for clustering tasks in the past are K-means and DBScan. We have first considered these algorithms for clustering the documents. Later, after doing some research we found out that strict clustering algorithms such as K-means will not be well suited for document clustering as K-means will give one label/category for a group which is not true in a real scenario as one document can be considered for multiple categories. One such algorithm which considers this is Latent Dirichlet Allocation (LDA) [1]. LDA gives you a probabilistic composition of the document - it produces a probability distribution of groupings (topics) per document. We also used HDBSCAN [13], a one-up on the DBSCAN, and uses a hierarchical clustering algorithm which is as the data is represented in multiple dimensions. Scatter/Gather, a clustering-based document browsing system for the textual domain, employed a hybrid strategy that included K-means [18] with Agglomerative hierarchical clustering.

The choice of dataset is explained later but we used the 20 newsgroup dataset. Due to its simplicity and high performance the bag of words model has always been preferred in literature and is one of the most widely used feature models for all kinds

of text classification tasks. The model represents the text to be classified as a bag or collection of individual words with no link of one word with the other, i.e. it completely disregards grammar and order of words within the text. This model is also very popular in sentiment analysis and has been used by various researchers. This is a very simplifying assumption but it has been shown to provide rather good performance. There are three ways of using the prior polarity of words as features. The simpler un-supervised approach is to use publicly available online lexicons/dictionaries which map a word to its prior polarity. Using publicly accessible internet lexicons or dictionaries that map a word to its preceding polarity is the easier unsupervised method.

IV. System Architecture and Algorithms

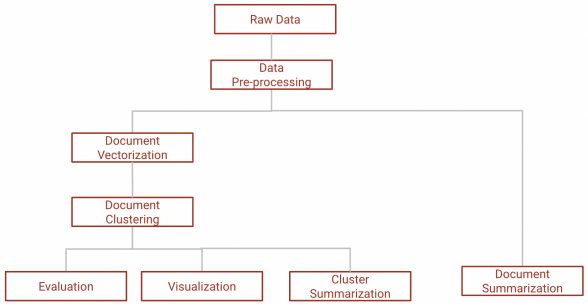

Fig. 1. System Architecture

A. Document Clustering

1) Latent Dirichlet Allocation: Latent Dirichlet Allocation [1] is a technique used to automatically categorize and understand lengthy documents. Documents are composed of a variety of words, and each topic has a variety of terms. LDA looks at the words in a document to identify the topics to which it belongs. LDA is similar to k-means clustering in some ways: both are unsupervised learning methods that don't require pre-trained data. However, while k-means clustering assigns a single topic label to each data set, LDA assigns multiple topic labels, which allows for a more nuanced understanding of the data.

2) HDBScan: HDBSCAN [13] is a hierarchical clustering algorithm that takes noise into account and is able to capture the shape or property of clusters, as well as their density. HDBSCAN is an implementation of the DBSCAN algorithm that allows for varying epsilon values, which means that it does not require a specific distance threshold to be chosen beforehand. HDBSCAN extends DBSCAN by converting it into a hierarchical clustering algorithm and then using a technique to extract a flat clustering based on the stability of clusters.

3) Agglomerative Clustering: Agglomerative hierarchical clustering [14] starts by considering every data point as its own cluster. Then, clusters are merged together to form a new cluster. This process is repeated until all the points in the data merge and form a single cluster. Agglomerative hierarchical clustering is a bottom-up approach, which means that at each iteration of the algorithm, larger clusters are formed. Unlike k-means clustering, you don't need to specify the number of clusters you want to divide the data into. Agglomerative hierarchical clustering can be used to find the appropriate number of clusters in a dataset.

B. Summarization

Text summarization can be done in two ways. The first one, extractive summarization, aims at identifying the most important sentences and using those exact sentences as the summary. In this method, first, we tokenize the words, and remove stop words and punctuation. Then we compute the frequency of each word in the document and calculate its weights. 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. Once the weights are assigned, the total weight for each sentence is calculated. The sentences with high weights are considered as important ones and appended to form the final summary. But the summary generated this way was not much meaningful so we moved to a more advanced approach, which is abstractive summarization.

In abstractive summarization, the meaning or the context of the document is taken into consideration and a new summary is generated. The sentences in the generated summary may be completely new sentences that are meaningful and accurately represent the document content. To implement this approach, Facebook BART Large CNN [19] has been used. The BART model is particularly effective when fine-tuned for text generation and can be used for text comprehension tasks. In the project, we chose an already trained BART model that was fine-tuned on CNN Daily Mail, a large collection of text-summary pairs. We loaded the pre-trained model and used its summarizer to generate a meaningful summary of the documents. Additionally, text from all the documents of a cluster is appended to form a large document, and summarization was done on it. The generated summary represents the summary of the entire cluster.

V. Datasets

A. Dataset Description

For this project, we explored a couple of publicly available datasets on Kaggle [5]. Initially, the Reuters dataset was explored. But, this dataset had multiple categories that are overlapping and non-exhaustive and there are relationships among the categories. Due to these reasons, the Reuters dataset was more suitable for a classification problem rather than a clustering problem. Later, a medium articles dataset was explored but the dataset has considerably less number of documents and we needed a larger dataset to perform clustering.

Finally, we moved on to the 20 newsgroups dataset [2]. This dataset is a collection of 20000 newsgroup documents. All these documents are evenly partitioned across 20 different newsgroups, each corresponding to a different topic. Additionally, the categories are not highly related as in the Reuters

dataset and has many documents making it the best fit for this problem statement.

B. Data Pre-processing and Vectorization

As part of data pre-processing, we first analyzed the 20 newsgroup dataset and took all the subsets of fetch20newsgroups. We have removed the headers and footers from the dataset and extracted the text and labels from the dataset into data frames. After that, we tokenized the sentences into tokens. Later, converted the tokens into lowercase alphabets and removed the punctuation, white spaces, stop words. Finally, applied lemmatization to normalize the sentence to keep the meaning of the sentence intact. Data is then converted to vector form and all the null rows are removed from it [7]. The final shape of our data consists of 18864 rows and 2 dimensions. This preprocessed data is later passed to google's universal sentence encoder to generate embedding of the text sentences which is later used by clustering algorithms. These embeddings can be used to reduce the amount of training data needed to achieve good clustering results, by effectively clustering the meanings of whole sentences rather than just individual words.

VI. Evaluations

Once the clustering is done, it is important to evaluate the clustering algorithm performance. For this, we have chosen five metrics that make the most sense for this problem statement

A. Homogeneity

This metric measures how much the sample in a cluster are similar. It ranges from 0 to 1. For example, when all samples that belong to a cluster 'k' are assigned the same label 'c', then the homogeneity would be 1.

B. Completeness

This metric measures how much similar samples are put together by the clustering algorithm. It ranges from 0 to 1. For example, when all the samples of with label 'c' are assigned to the same cluster 'k' then completeness would be 1.

C. Adjusted Rand Index

This metric determines if two clusters are similar to each other. It ranges from 0 to 1. 0 adjusted rand index means random labelling and 1 means the clusters are identically partitioned.

D. V-measure

V-measure measures the goodness of clustering partition. It is the harmonic average of homogeneity and completeness. If either of homogeneity or completeness is low, then the V-measure will also be low.

E. Silhouette Score

This metric is used to calculate the goodness of a 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.

The above five evaluation metrics have been calculated for all the clustering techniques mentioned in the algorithms section as shown in Table 1. We noticed that LDA performed better against the HDBScan and Agglomerative Clustering in four of the five metrics - Homogeneity, Completeness, V-measure, and Adjusted Rand Index. Whereas HDBScan performed better in terms of the Silhouette score evaluation metric. Finally, we infer that LDA was the best clustering algorithm suited 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 visualization of high-dimensional datasets is especially well suited to the dimensionality reduction method known as t-Distributed Stochastic Neighbor Embedding (t-SNE) [16]. The data is shown in a lower dimension via t-SNE while maintaining its local structure. In order to maximize the similarity measures, a cost function is used to compute a similarity measure between pairs of instances in both the high-dimensional and low-dimensional space.

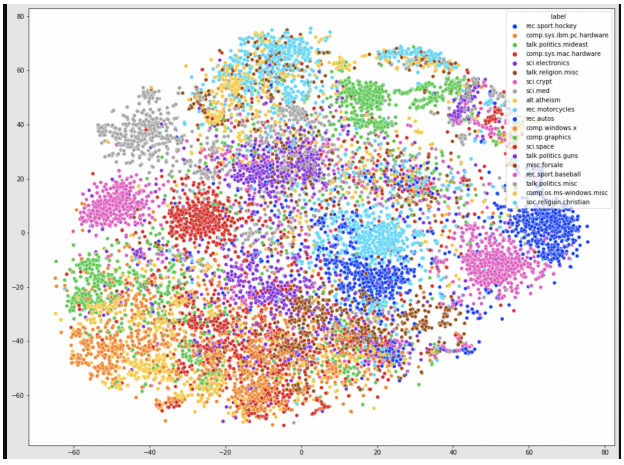

Fig. 2. Visualization of LDA clusters using t-SNE

B. UMAP

A dimension reduction method called Uniform Manifold Approximation and Projection (UMAP) [8] can be utilized for generic non-linear dimension reduction as well as visualization in a manner similar to t-SNE. UMAP uses several

TABLE I
EVALUATION RESULTS OF CLUSTERING ALGORITHMS

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

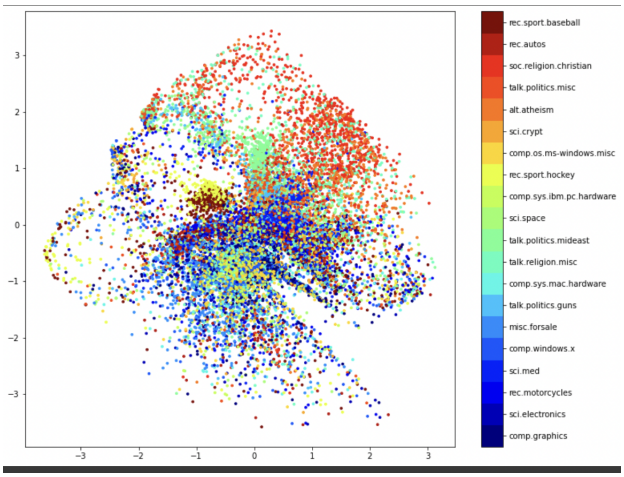

Fig. 3. Visualization of HDBScan clusters using t-SNE

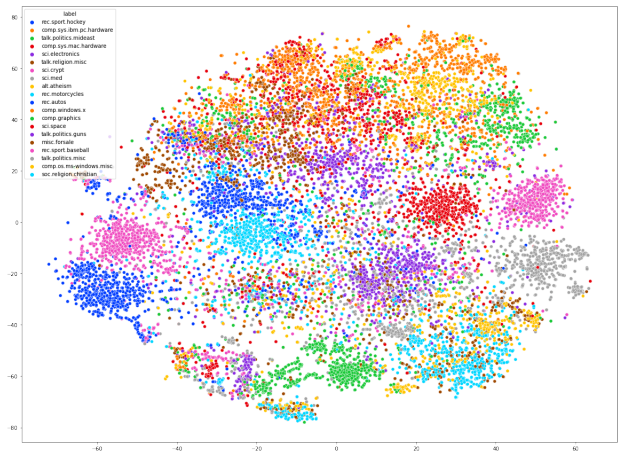

Fig. 4. Visualization of Agglomerative clusters using t-SNE

optimization approaches to speed up the process, making it faster than t-SNE. Without dimensionality reduction and data pre-processing, UMAP can be used on the dataset.

C. CVAE

Compression The Variational Autoencoder (VAE) [9] is a dimensionality reduction technique that expands on the simplicity of the t-SNE and UMAP implementations while introducing a number of very desirable features. It is quicker than t-SNE or UMAP because it is based on variational autoencoders. Although it requires a large amount of data to be trained, CVAE scales well to high dimensional input and latent spaces in contrast to t-SNE and UMAP, which perform

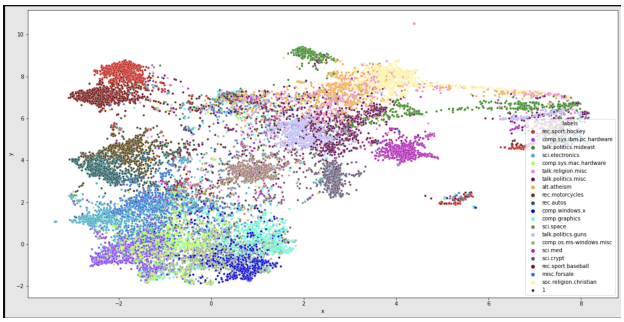

Fig. 5. Visualization of LDA clusters using UMAP

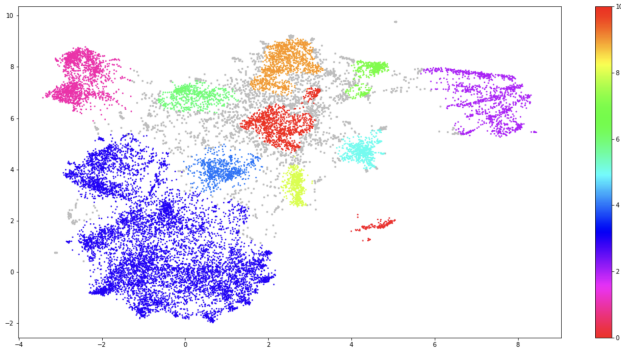

Fig. 6. Visualization of HDBScan clusters using UMAP

better even with smaller sets of data. Additionally, CVAE often obtains a weak separation between clusters.

VIII. TEAM CONTRIBUTION

The contribution of the team members can be outlined below

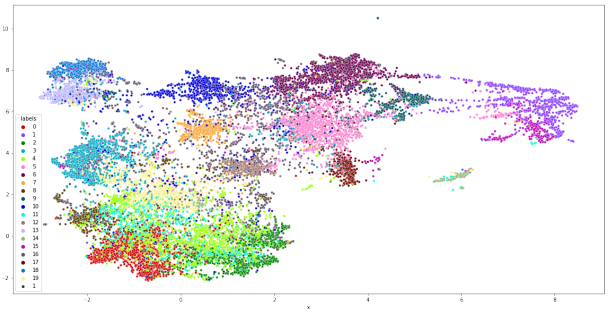

Fig. 7. Visualization of Agglomerative clusters using UMAP

TABLE II
CONTRIBUTIONS OF EACH TEAM MEMBERS

Task	Team Members	Deadline
Study of clustering and visualization techniques	All team members	Aug 31 - Sep 16
Data Pre-processing	Swaapnika, Harshitha, Mounika	Sep 17 - Sep 28
Data Embedding	Yasir, Akash, Sagar	Sep 28 - Oct 2
Clustering	Akash, Sagar, Yasir, Mounika	Oct 2 - Oct 30
Document summarization	All team members	Nov 1 - Nov 14
Visualization	Swaapnika, Harshitha, Sagar, Yasir	Nov 5 - Nov 18
Evaluation and analysis	Mounika, Swaapnika, Harshitha, Akash	Nov 19 – Nov 22
Summary and final evaluation	All team members	Nov 23 - Dec 2

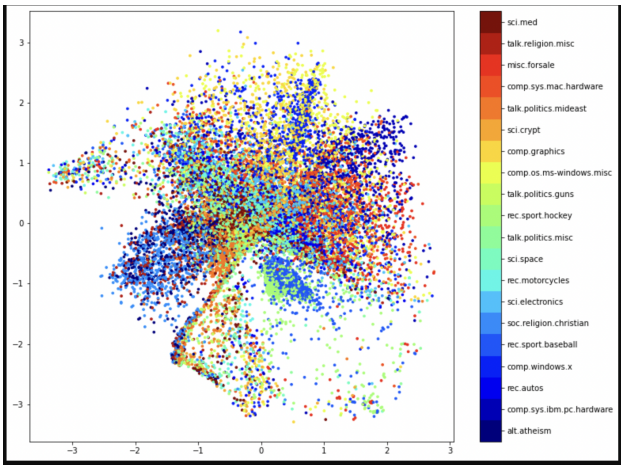

Fig. 8. Visualization of LDA clusters using CVAE

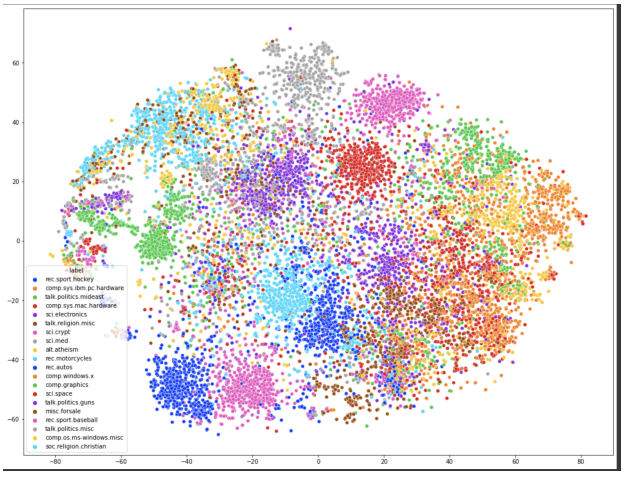

Fig. 9. Visualization of HDBScan clusters using CVAE

A. Study of clustering and visualization techniques

Perform research on different clustering and visualization techniques to apply on datasets. All the team members have researched the algorithms to be used.

B. Data Pre-processing

Data pre-processing is handled by Swaapnika, Harshitha, Mounika. The preprocessing was done by removing the noise from the data and converting it to process for data embedding.

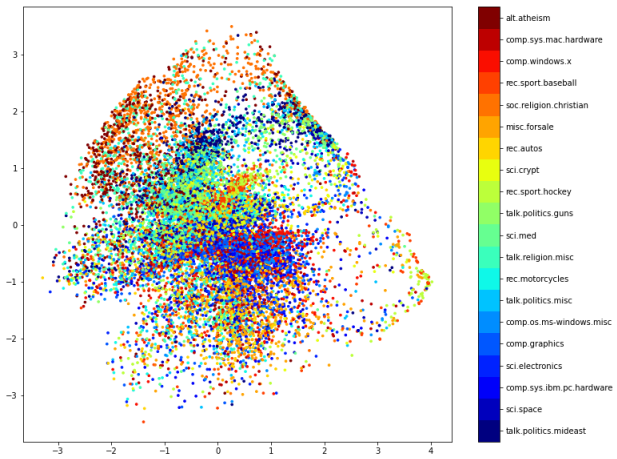

Fig. 10. Visualization of Agglomerative clusters using CVAE

Cluster: rec.sport.hockey
Original Document:
Doc1: I am sure some bashers of Pens fans are pretty confused about the lack of any kind of posts about the recent Pens massacre of the Devils. Actually, I am bit puzzled too and a bit relieved ……..
Doc2: Ottawa picks #1 which means it is almost 100% that Alexander Daigle will go #1. He'll either stay or be traded in Montreal or Quebec. IMO I would take Kariya. He should alot of leadership in the NCAA and so far in the World Championships. Daigle didn't show this for his junior team.
San Jose will then get Kariya…..
.
.
.
Other Documents

Cluster Summary:
The Pens are killing those Devils worse than I thought. Jagr just showed you why he is much better than his regular season stats. Bowman should let JAgr have a lot of fun in the next couple of games. I was very disappointed not to see the Islanders lose the final regular season game.Alexander Daigle is almost 100% that Ottawa picks #1. San Jose will then get Kariya. Tampa Bay will either go for a russian Kozlov (I think that's it) or a defenseman Rob Niedemeyer.Here are the NHL's alltime leaders in goals and points at the end of the 1992-3 season…..

Fig. 11. Output of document summarization of a cluster

C. Data Embedding

Data embedding is implemented by Yasir, Akash and Sagar. The preprocessed data was converted to vector form using google's universal sentence encoder to generate sentence encoding.

D. Clustering

We have explored multiple clustering algorithms to cluster the documents. We have then finalized with LDA, Agglomerative clustering, HDBScan and have split the implementation among Akash, Sagar, Yasir and Mounika.

E. Document Summarization

We have implemented two ways of summarizing using Abstractive and Extractive text summarization techniques. All the team members have contributed to the summarization part.

F. Document Visualization

The module to implement the document visualization was using UMAP, T-SNE, Compression VAE was handled by Swaapnika, Harshitha, Sagar and Yasir.

G. Evaluation and Analysis

The task of figuring out suitable evaluation metrics for document clustering and analysing the results were handled by Mounika, Swaapnika, Harshitha and Akash. We have finalized with Homogeinity, Completeness, V-measure, Adjusted Rand-Index, Silhouette Coefficient evaluation metrics.

H. Summary and Documentation

All the team members have contributed to the project report by documenting some sections of the report.

IX. Conclusion

In conclusion, throughout this research, we conducted a subjective study through the visualization of the clusters, which enabled us to carefully examine the dataset. In order to do clustering, we were able to use techniques like LDA, HDBScan, and K-Means. UMAP and t-SNE were then used to visualize the clustered categories. Utilizing the assessment metrics we established, a comparison of the algorithms stated was conducted. It was clear that LDA was producing the best results possible with the available dataset.

We did extractive text summarization [10] using SpaCy [20]. We loaded the model and assign weights based on word frequency. Then we are using Facebook's BART large CNN to take that vector to create a summarization using a sentence token. Then using the cluster in the data frame we summarize the sentence in less than 2700 chars for the document result summarization. The Facebook BART is a transformer encoder-decoder. Having BERT and GPT3 features, somewhat similar or even better to the T5 model. Thus the summarization is State of the art and we tested that it gave a relevant summary of the topics. In addition to being able to capture the sense of related categories in the dataset.

References

[1] Giri. (2021, May 2). Is Latent Dirichlet Allocation (LDA) A clustering algorithm? HDS; High Demand Skills. https://highdemandskills.com/lda-clustering/.

[2] http://qwone.com/ jason/20Newsgroups/.

[3] https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection.

[4] https://albertauyeung.github.io/2020/06/19/bert-tokenization.html/.

[5] Karmakar, Saurav. "Syntactic and Semantic Analysis and Visualization of Unstructured English Texts." (2011).

[6] Millar, Jeremy R. et al. "Document Clustering and Visualization with Latent Dirichlet Allocation and Self-Organizing Maps." FLAIRS Conference (2009).

[7] Cao, Tuan-Dung et al. "Hot Topic Detection on Newspaper" Conference: the Ninth International Symposium (2018)

[8] arXiv:2012.04456 "Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap,and PaCMAP for Data Visualization".

[9] Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap,and PaCMAP for Data Visualization

[10] 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

[11] 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.

[12] https://towardsdatascience.com/compressionvae-a-powerful-and-versatile-alternative-to-t-sne-and-umap-5c50898b8696.

[13] HDBScan - https://towardsdatascience.com/tuning-with-hdbscan-149865ac2970.

[14] Agglomerative-hierarchical-clustering - https://medium.com/geekculture/agglomerative-hierarchical-clustering-a-gentle-intro-with-an-example-program-4b7afe35fd4b.

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

[16] https://lvdmaaten.github.io/tsne/.

[17] https://towardsdatascience.com/compressionvae-a-powerful-and-versatile-alternative-to-t-sne-and-umap-5c50898b8696.

[18] Lloyd, S. P. (1957). Least squares quantization in PCM. Technical Report RR-5497, Bell Lab, September 1957.

[19] BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

[20] https://aparnamishra144.medium.com/automated-text-summarization-using-spacy-in-nlp-8750b1b6e404