# Cybersecurity Policy Clustering with LLM-Based Embeddings and Dimensionality Reduction

Adriana García Aguirre[1], Pablo Rivas[2], and Liang Sun[1]

[1] Department of Mechanical Engineering, Baylor University, Texas, USA
{Adriana_GarciaAguir1,Liang_Sun}@Baylor.edu
[2] Department of Computer Science, Baylor University, Texas, USA
Pablo_Rivas@Baylor.edu

**Abstract.** The clustering of cybersecurity policy documents presents a significant challenge in legal Natural Language Processing (NLP), particularly within government and defense sectors. This study evaluates the effectiveness of clustering techniques when applied to cybersecurity policies represented using BERT-based embeddings. We employ dimensionality reduction methods, including Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP), to project high-dimensional embeddings into lower-dimensional space. We then assess the performance of K-Means, DBSCAN, and Hierarchical Clustering in organizing policy documents. Our results indicate that UMAP combined with Hierarchical Clustering achieves the highest clustering performance, attaining an Adjusted Rand Index (ARI) of 0.8529. These findings demonstrate the impact of transformer-based language models on cybersecurity policy analysis and highlight the role of dimensionality reduction in improving clustering effectiveness.

**Keywords:** Cybersecurity Policy Clustering · Transformer Models · Natural Language Processing · Dimensionality Reduction

## 1 Introduction

The increasing sophistication and frequency of cyber threats have underscored the critical need for advanced analytical techniques to enhance cybersecurity defenses. As artificial intelligence (AI) and machine learning (ML) continue to play a central role in cybersecurity, the availability of structured and high-quality datasets has become indispensable for training robust models. In particular, NLP techniques have demonstrated their effectiveness in extracting actionable insights from policy documents, threat reports, and security logs. However, challenges remain in classifying and organizing cybersecurity policies due to the inherent complexity and overlap among policy documents. This study builds upon prior research by Quevedo et al. [14], who developed a dataset for natural language understanding (NLU) in the context of Department of Defense (DoD) cybersecurity policies. Our research extends this work by leveraging advanced clustering and

visualization techniques to analyze the latent structure of cybersecurity policy documents.

Effective policy classification requires methods that can capture subtle semantic differences and structural patterns within textual data. As highlighted by Mahn et al. [10], integrating NLP and ML techniques is essential for improving cybersecurity frameworks, particularly in policy interpretation and compliance assessment. Clustering algorithms such as K-Means and hierarchical clustering have proven useful in organizing and analyzing large-scale textual datasets, as demonstrated by Probierz et al. [7]. By applying these techniques to the DoD cybersecurity dataset, this research aims to uncover meaningful policy clusters, aiding in the refinement of classification strategies and improving interpretability.

Dimensionality reduction plays a crucial role in facilitating the visualization of high-dimensional data, making it possible to identify patterns and overlaps in cybersecurity policy clusters. Techniques such as t-SNE and UMAP have been widely adopted for this purpose, preserving both local and global structures within the data [12]. The effectiveness of such methods has been demonstrated in prior work by George and Sumathy, who integrated clustering and NLP-based approaches to enhance topic modeling outcomes [5]. By employing these techniques, our study aims to provide an interpretable representation of DoD cybersecurity policies, enabling a more nuanced understanding of policy classifications and potential areas of ambiguity.

This research contributes to the broader discussion on AI-driven policy analysis by demonstrating how visualization and clustering techniques can improve the classification and interpretation of cybersecurity documents. The insights gained from this study are expected to inform both academic research and practical cybersecurity applications, facilitating better policy management and compliance monitoring. By systematically identifying overlapping clusters and refining classification approaches, our work lays the foundation for more sophisticated AI-driven methodologies in cybersecurity policy analysis.

## 2 Background

The foundation of this research builds upon the work of Quevedo et al., who developed the CSIAC-DoDIN V1.0 dataset for NLU within the context of DoD cybersecurity policies [14,2]. Their work introduced a structured dataset extracted from cybersecurity policy documents, incorporating key attributes such as classification, purpose, scope, applicability, type, and textual content. This dataset serves as a critical resource for training machine learning models in cybersecurity-related NLP tasks.

The application of clustering techniques in textual datasets has been widely explored in prior research. Probierz et al. proposed a method for clustering scientific literature based on thematic content, employing NLP and the K-Means algorithm to categorize academic articles [13]. Their methodology involved preprocessing text by standardizing case formats, removing non-alphabetic characters,

applying tokenization, normalization (stemming and lemmatization), and filtering stopwords. The processed text was then transformed into token count matrices using binary measures, term frequency (TF), and term frequency-inverse document frequency (TF-IDF), which served as the foundation for clustering. Their evaluation of clustering effectiveness was conducted using a connection matrix derived from shared keywords, with experiments spanning 1,557 articles published between 2017 and 2022. Their findings indicate that TF-IDF produced the most thematically cohesive clusters, demonstrating superior connection coefficients compared to alternative text representation methods. While increasing the number of clusters enhanced thematic specificity, it simultaneously reduced generalization, highlighting a trade-off between precision and broad applicability.

Advancements in domain-specific NLP models have further improved text analysis capabilities in cybersecurity contexts. Bayer et al. introduced CySecBERT, a cybersecurity-focused language model based on BERT, specifically trained on a diverse corpus encompassing cybersecurity blogs, scientific literature, and social media content [1]. Designed to mitigate catastrophic forgetting, the model's training process carefully balanced hyperparameters such as learning rate, dataset size, and the number of training epochs. The model was evaluated through both intrinsic and extrinsic tasks, demonstrating significant improvements in classification and named entity recognition (NER), surpassing existing models such as CyBERT. Notably, CySecBERT proved particularly effective for domain-specific tasks, including cyber threat intelligence (CTI) analysis. However, its general NLP performance was lower than that of standard BERT, which is expected given its specialized training focus. To foster further research and development, CySecBERT and its dataset have been made publicly available, with potential applications in phishing detection, malware analysis, and cybersecurity threat identification. The authors, however, caution against inherent social biases present in the dataset and emphasize the need for ethical and responsible AI deployment in cybersecurity applications.

The application of clustering techniques to unstructured text data has been extensively explored in various domains, including aviation safety and biomedical research. Rose et al. developed an NLP-based method for clustering and analyzing aviation safety reports collected by the Aviation Safety Reporting System (ASRS) [15]. This system compiles voluntary and anonymous safety incident reports from pilots, air traffic controllers, and maintenance personnel, providing a valuable resource for identifying patterns in safety-related events. Their study focused on passenger and cargo operations between 2010 and 2020, selecting 13,336 reports for analysis. To uncover meaningful structures in the data, the authors employed K-Means clustering in conjunction with t-SNE, a dimensionality reduction technique. Their approach identified ten primary groups and 31 subgroups, revealing latent patterns in aviation safety narratives. PCA was further utilized to reduce the feature space from 1,000 to 150 dimensions, demonstrating the efficacy of clustering for extracting trends and potential risk factors from large-scale textual datasets.

In the context of topic modeling, George and Sumathy proposed a novel framework integrating BERT with Latent Dirichlet Allocation (LDA) and K-Means clustering to improve the extraction of thematic patterns from large-scale unstructured text corpora [5]. Their methodology was applied to the CORD-19 dataset, a comprehensive collection of COVID-19 research papers, preprints, and metadata. The preprocessing pipeline included data cleaning, integration, and transformation steps, followed by the application of TF-IDF to assess word importance within the dataset. To enhance the representation of textual content, the researchers employed multiple dimensionality reduction techniques, including PCA, t-SNE, and UMAP. Their framework utilized BERT for generating contextualized sentence embeddings, while LDA was leveraged to identify latent topics. K-Means clustering was then applied to group topics into coherent thematic structures. The effectiveness of the clustering approach was evaluated using the Silhouette Score, which quantified the quality of the generated clusters. Their findings indicate that the integration of clustering with dimensionality reduction techniques enhances the interpretability and coherence of discovered topics, demonstrating the potential for applying similar methodologies to cybersecurity policy analysis.

## 3 Methodology

This study aims to analyze the structure of cybersecurity policy documents by leveraging deep learning-based text embeddings, dimensionality reduction techniques, and clustering algorithms. The methodology builds upon the CSIAC-DoDIN V1.0 dataset [14,2], which comprises structured data extracted from DoD cybersecurity policies. The objective is to project these high-dimensional representations into a lower-dimensional space, allowing for visualization and evaluation of clustering effectiveness. The proposed workflow, illustrated in Fig. 1, consists of four primary steps: document embedding using BERT, dimensionality reduction, clustering, and visual analysis.

### 3.1 Text Embedding Using BERT

The preprocessing phase involves converting cybersecurity policy documents into high-dimensional numerical representations using BERT (Bidirectional Encoder Representations from Transformers). BERT is a transformer-based language model that generates contextualized embeddings, capturing both syntactic and semantic relationships between words [3]. Given a document represented as a sequence of tokens, the BERT model maps each token to a dense vector representation in a high-dimensional latent space. The resulting embeddings are obtained by averaging the token representations across the document, producing a fixed-size vector $\mathbf{E} \in \mathbb{R}^d$, where $d = 768$ for BERT-base and $d = 1024$ for BERT-large. These embeddings serve as input for subsequent dimensionality reduction techniques.

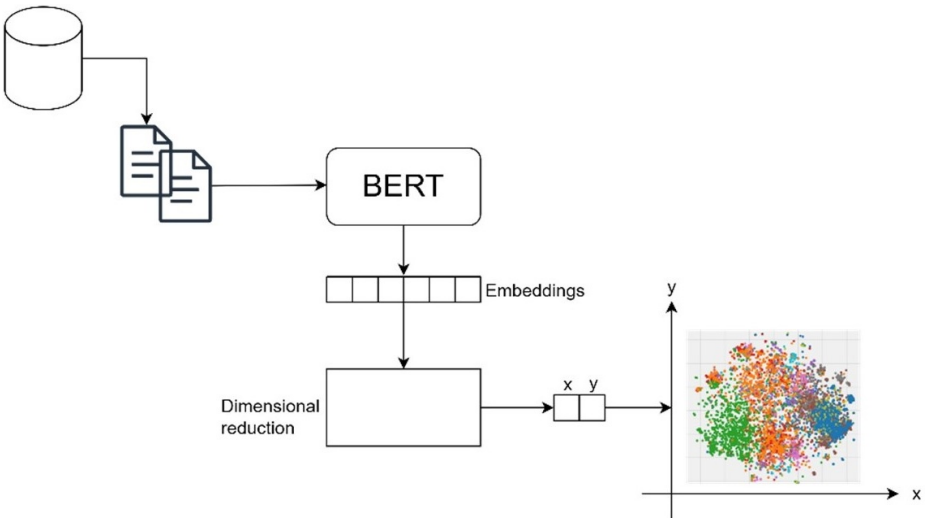


**Fig. 1.** Overview of the methodology: Documents are processed through BERT to generate embeddings, which are then reduced to two dimensions using PCA, t-SNE, or UMAP. The resulting data points are clustered and visualized for analysis.

### 3.2 Dimensionality Reduction

To facilitate visualization and clustering, the high-dimensional document embeddings must be projected into a two-dimensional space while preserving the underlying structure. Three dimensionality reduction techniques were evaluated: PCA, t-SNE, and UMAP.

PCA is a linear transformation technique that identifies orthogonal axes, or principal components, that maximize the variance in the dataset [16]. Given a set of embeddings $\mathbf{X} \in \mathbb{R}^{n \times d}$, PCA seeks to find a lower-dimensional representation $\mathbf{Z} \in \mathbb{R}^{n \times m}$ such that $\mathbf{Z} = \mathbf{XW}$, where $\mathbf{W}$ contains the top $m$ eigenvectors of the covariance matrix $\mathbf{X}^T\mathbf{X}$. While PCA is computationally efficient, it assumes linear relationships among features, which may not always be the case in complex NLP tasks.

In contrast, t-SNE is a non-linear technique that models pairwise similarities in high-dimensional space and maps them onto a lower-dimensional manifold [9]. It minimizes the Kullback–Leibler (KL) divergence between probability distributions in the original and reduced spaces, preserving local neighborhood structures. Although t-SNE is effective in revealing fine-grained cluster structures, it is computationally intensive and does not always maintain global data relationships.

UMAP offers an alternative non-linear approach that constructs a high-dimensional graph representation of the data before optimizing its projection into a lower-dimensional space [11]. Unlike t-SNE, UMAP balances local and global structure preservation while maintaining greater scalability for large datasets.

By leveraging a topological framework based on Riemannian geometry, UMAP provides a computationally efficient solution for high-dimensional text embeddings.

### 3.3 Clustering Algorithms

Once the data is projected into a two-dimensional space, clustering techniques are employed to identify meaningful patterns in cybersecurity policies. Three clustering algorithms were considered: K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and hierarchical clustering.

K-Means is a partition-based clustering method that minimizes the within-cluster variance by iteratively updating cluster centroids [8]. Given a set of data points $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$, the algorithm partitions them into $k$ clusters by solving the optimization problem:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 \tag{1}$$

where $\mu_i$ represents the centroid of cluster $S_i$. K-Means assumes that clusters are convex and isotropic, which may not hold in real-world datasets with irregular cluster shapes.

To address this limitation, DBSCAN is employed as a density-based clustering method that groups points based on their spatial density, allowing the discovery of arbitrarily shaped clusters [4]. Unlike K-Means, DBSCAN does not require a predefined number of clusters and effectively identifies outliers as noise. However, its performance is highly dependent on the selection of hyperparameters, particularly the neighborhood radius $\varepsilon$ and minimum points per cluster.

Finally, hierarchical clustering constructs a tree-like structure (dendrogram) by iteratively merging or splitting clusters based on a distance metric [6]. This method provides a hierarchical representation of the data, enabling flexibility in selecting the optimal number of clusters. However, it is computationally more expensive than K-Means and DBSCAN, making it less suitable for large-scale datasets.

By integrating BERT-based embeddings, dimensionality reduction, and clustering, this study aims to improve the interpretability of cybersecurity policy classifications and provide a structured analysis of policy document distributions in latent space.

## 4 Experiments

### 4.1 Dataset Description

The dataset used in this study consists of 7,698 documents, encompassing cybersecurity-related policies, guidelines, strategies, responsibilities, and procedures issued by the DoD. These documents have been systematically categorized based on their

intended policy objectives. The dataset is structured into hierarchical clusters, each corresponding to a distinct aspect of cybersecurity governance, including policy organization, access control, attack prevention, and operational strategy. Fig. 2 illustrates the hierarchical clustering of these policy documents, showing the relationship between various policy domains.

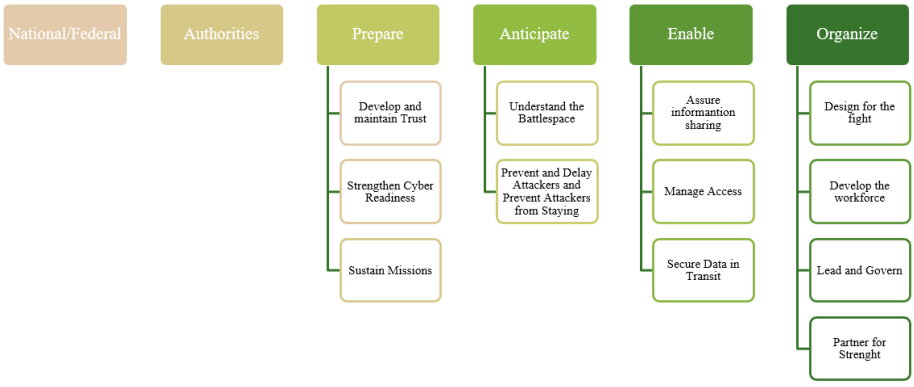


**Fig. 2.** Visualization of DoD cybersecurity policy clusters based on hierarchical classification.

Each document in the dataset is annotated with several attributes, including document ID, cluster assignment, classification, source, purpose, scope and applicability, type (policy, responsibility, or procedure), and full text. These annotations enable multiple downstream tasks, such as cluster classification, subcluster classification, and text entailment analysis. Notably, one identified limitation is the absence of text examples in the "Lead and Govern" subcluster, as these documents primarily pertain to strategic directives rather than operational policies.

### 4.2 Data Visualization

To gain insights into the distribution and structure of the dataset, the raw document embeddings obtained from BERT were projected into a lower-dimensional space using three different dimensionality reduction techniques: PCA, t-SNE, and UMAP. The goal of this visualization step is to analyze how well the policy clusters are separated in reduced space before applying clustering algorithms.

The preprocessing workflow involved importing the dataset, applying tokenization and embedding extraction using the BERT-base-uncased model, and then performing dimensionality reduction. Fig. 3 presents an overview of the original data projected into two dimensions using these techniques. The results demonstrate varying levels of cluster separability, with UMAP showing the most distinct grouping.

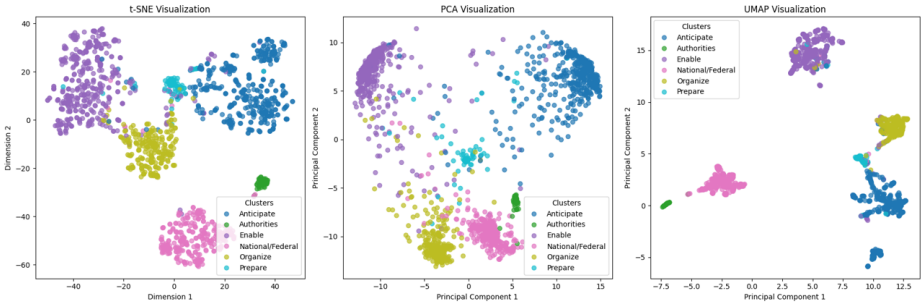


**Fig. 3.** Comparison of different dimensionality reduction techniques applied to DoD cybersecurity policy embeddings.

### 4.3 Clustering Analysis

To evaluate the effectiveness of different clustering methods in structuring cybersecurity policies, we applied three widely used clustering algorithms: K-Means, DBSCAN, and Hierarchical Clustering. Each algorithm was tested using three different dimensionality reduction techniques: PCA, t-SNE, and UMAP.

Fig. 4 presents the clustering results obtained using PCA as the dimensionality reduction method. The K-Means algorithm produced distinct clusters with well-defined boundaries, while DBSCAN struggled with noisy data points. Hierarchical clustering demonstrated a structured separation of policy categories but exhibited some overlap in certain regions.

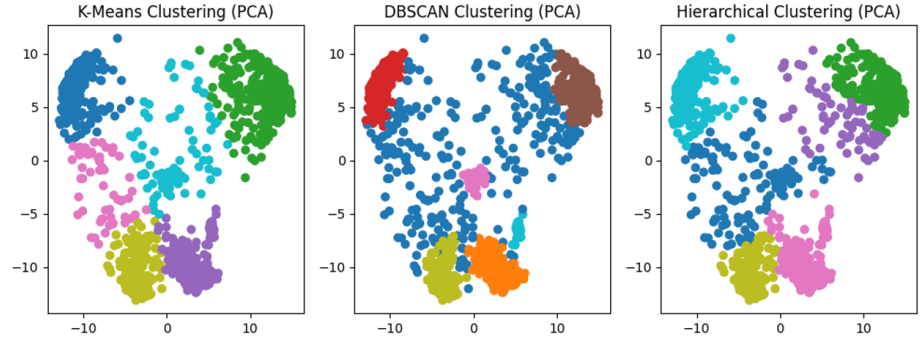


**Fig. 4.** Clustering results using PCA for dimensionality reduction.

Fig. 5 shows the clustering results obtained using t-SNE. While t-SNE provided a clearer separation of clusters than PCA, the K-Means algorithm still suffered from some misclassification of policy groups. DBSCAN exhibited difficulties in identifying clear boundaries, and hierarchical clustering demonstrated improved performance compared to PCA-based clustering.

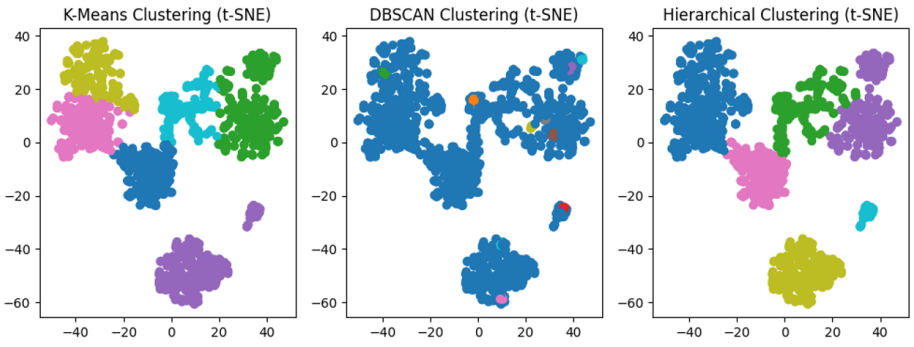


**Fig. 5.** Clustering results using t-SNE for dimensionality reduction.

Finally, Fig. 6 illustrates the clustering performance using UMAP. Among the three dimensionality reduction techniques, UMAP exhibited the most compact and well-separated clusters. The K-Means algorithm effectively identified major policy categories, while DBSCAN and hierarchical clustering provided robust alternative groupings, particularly in detecting outlier documents and subclusters.

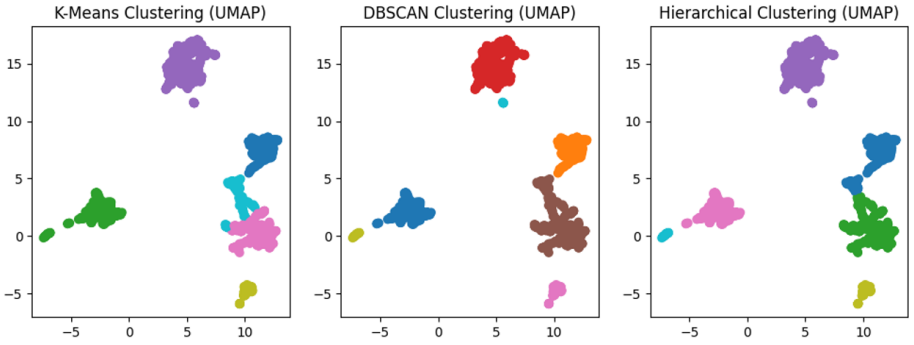


**Fig. 6.** Clustering results using UMAP for dimensionality reduction.

Overall, the experimental results demonstrate that UMAP, combined with hierarchical clustering, achieves the best separation of policy clusters, highlighting its potential for analyzing complex legal and cybersecurity-related textual data. The findings suggest that document embedding methods, when coupled with appropriate clustering techniques, can enhance the interpretability and organization of cybersecurity policies, aiding in policy classification and retrieval.

# 5 Analysis

## 5.1 Clustering Performance with PCA

To evaluate the effectiveness of clustering algorithms, we first applied PCA for dimensionality reduction before clustering the transformed data. Fig. 7 presents the results obtained using K-Means, DBSCAN, and Hierarchical Clustering. The plots indicate that both K-Means and Hierarchical Clustering successfully capture the primary cluster structures, though some overlap is still evident. DBSCAN, while effective at identifying dense cluster centers, leaves a significant number of points unclustered due to its sensitivity to density parameters.

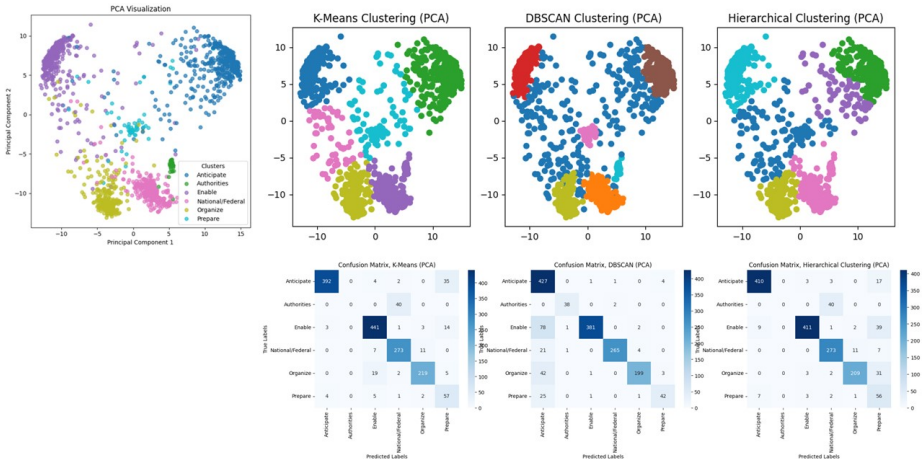


**Fig. 7.** Comparison of clustering results using PCA for dimensionality reduction. The confusion matrices illustrate how well each clustering algorithm assigns data points to the predefined policy categories.

The confusion matrices derived from the cluster assignments indicate that the overall clustering quality is acceptable. K-Means and Hierarchical Clustering demonstrate reasonable classification accuracy, though some misclassification occurs due to overlapping clusters. DBSCAN struggles with sparsely distributed policies, leading to a higher number of unassigned points.

## 5.2 Clustering Performance with t-SNE

Applying t-SNE for dimensionality reduction yielded improved separation between clusters, as shown in Fig. 8. Hierarchical Clustering performed best under this setup, correctly separating clusters while maintaining structural integrity.

K-Means demonstrated some misclassification, particularly in the upper-right cluster, where it split a single group into two. Additionally, two smaller clusters at the bottom were incorrectly merged. DBSCAN, even after extensive parameter tuning, struggled to identify well-defined clusters, highlighting its limitations when dealing with variable-density distributions.

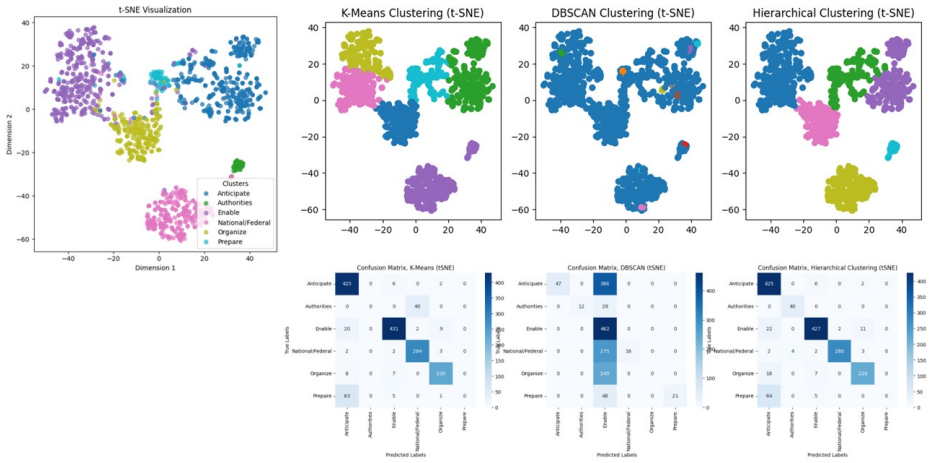


**Fig. 8.** Comparison of clustering results using t-SNE for dimensionality reduction. The confusion matrices reveal how each method handles high-dimensional data when projected into a lower-dimensional space.

The confusion matrices illustrate that smaller clusters are more prone to being absorbed into larger clusters in both K-Means and Hierarchical Clustering. This effect is less pronounced in t-SNE than in PCA, suggesting that t-SNE provides a more meaningful representation of the document embeddings.

### 5.3 Clustering Performance with UMAP

UMAP produced the most distinct cluster separations among the three dimensionality reduction techniques, as shown in Fig. 9. UMAP preserves both global and local structures effectively, leading to well-separated clusters in the projected space. As a result, all three clustering algorithms—K-Means, DBSCAN, and Hierarchical Clustering—achieved superior performance when applied to the UMAP-reduced data.

Hierarchical Clustering demonstrated the best overall performance, capturing the natural structure of the dataset while minimizing misclassifications. K-Means also produced strong results but showed some inconsistencies in handling smaller clusters. Interestingly, DBSCAN exhibited significant improvements compared to

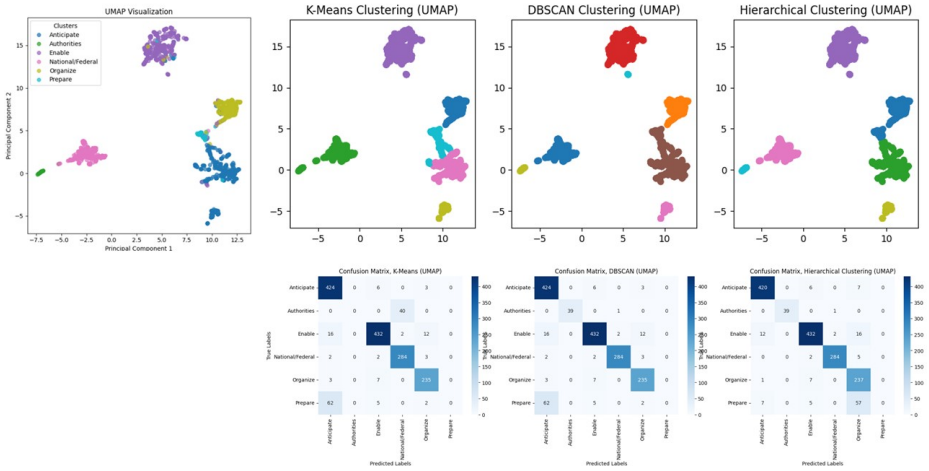


**Fig. 9.** Comparison of clustering results using UMAP for dimensionality reduction. UMAP provides better-separated clusters, resulting in improved clustering performance.

its performance with PCA and t-SNE, likely due to UMAP's ability to maintain cluster compactness.

### 5.4 Quantitative Evaluation

To quantify clustering performance, we computed the ARI for each combination of dimensionality reduction and clustering technique. The results, shown in Table 1, reveal that UMAP, when paired with Hierarchical Clustering, achieved the highest ARI score (0.8529), indicating strong alignment with ground-truth labels. K-Means also performed well across all three dimensionality reduction techniques, with ARI scores consistently above 0.80. In contrast, DBSCAN struggled with PCA and t-SNE, but demonstrated noticeable improvements when applied to UMAP-reduced data.

**Table 1.** ARI Scores for Different Clustering Methods and Dimensionality Reduction Techniques

| **Method** | **PCA** | **t-SNE** | **UMAP** | **Avg** |
|---|---|---|---|---|
| K-Means | **0.8205** | 0.8007 | 0.8108 | 0.8107 |
| DBSCAN | 0.6928 | 0.0216 | **0.8336** | 0.5160 |
| Hierarchical | 0.8096 | 0.7981 | **0.8529** | **0.8202** |
| **Avg** | 0.7743 | 0.5401 | **0.8324** | |

Overall, the experimental results indicate that combining UMAP with Hierarchical Clustering yields the most reliable clustering performance. This combination provides a balance between global cluster structure and local density preservation, making it well-suited for cybersecurity policy classification. K-Means also offers strong performance but exhibits some sensitivity to initial centroid selection. DBSCAN, while generally less effective, benefits significantly from the enhanced structure preservation provided by UMAP.

These findings underscore the importance of selecting an appropriate dimensionality reduction technique when clustering high-dimensional text data. Future work will explore optimizing hyperparameters for each clustering method and investigating hybrid approaches to further enhance clustering accuracy.

## 6 Conclusions

This study evaluated the effectiveness of different dimensionality reduction techniques and clustering algorithms for organizing and analyzing cybersecurity policy documents. Three dimensionality reduction methods, PCA, t-SNE, and UMAP, were assessed, with UMAP demonstrating the best performance in preserving cluster separability. Among the clustering algorithms tested, K-Means, DBSCAN, and Hierarchical Clustering, Hierarchical Clustering consistently outperformed the others in terms of accuracy and alignment with the ground truth.

The highest clustering performance was achieved by the combination of UMAP for dimensionality reduction and Hierarchical Clustering, which attained an ARI of 0.8529. This result indicates that UMAP effectively preserves both local and global structures in the data, enabling Hierarchical Clustering to identify well-separated and coherent clusters. In contrast, DBSCAN exhibited inconsistent performance, particularly when paired with PCA and t-SNE, likely due to its sensitivity to density variations within the dataset.

These findings suggest that the choice of dimensionality reduction technique significantly impacts the effectiveness of clustering methods when dealing with high-dimensional text data. Future work will explore multi-step dimensionality reduction approaches, where embeddings are progressively reduced from high-dimensional space (e.g., 512 to 256 dimensions) before reaching the final 2D representation. Additionally, alternative clustering strategies, such as hybrid approaches that combine the strengths of multiple algorithms, will be investigated to further improve classification accuracy and robustness.

**Acknowledgments.** The authors thank the Rivas.AI Lab (`https://lab.rivas.ai`) for their support and helpful feedback throughout this project. This work was funded in part by the National Science Foundation under grants CNS-2210091 and CNS-2136961, and by the U.S. Department of Education under grant P116Z230151.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References


1. Bayer, M., Kuehn, P., Shanehsaz, R., Reuter, C.: Cysecbert: A domain-adapted language model for the cybersecurity domain. Acm Transactions on Privacy and Security **27**, 1–20 (2024). https://doi.org/10.1145/3652594
2. Caballero, E.Q., Rivas, P., Arguelles, A.P., Rodriguez, A., Yero, J., Pienta, D., Cerny, T.: Natural Language Understanding Dataset for DoD Cybersecurity Policies (CSIAC-DoDIN V1.0) (11 2023). https://doi.org/10.6084/m9.figshare.22800185.v2
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4171–4186 (2019). https://doi.org/10.18653/v1/N19-1423
4. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. pp. 226–231 (1996)
5. George, L., Sumathy, P.: An integrated clustering and bert framework for improved topic modeling. International Journal of Information Technology **15**, 2187–2195 (2023). https://doi.org/10.1007/s41870-023-01268-w
6. Jr., J.H.W.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association **58**(301), 236–244 (1963). https://doi.org/10.1080/01621459.1963.10500845
7. Kam, H., Katerattanakul, P.: Enhancing student learning in cybersecurity education using an out-of-class learning approach. Journal of Information Technology Education Innovations in Practice **18**, 029–047 (2019). https://doi.org/10.28945/4200
8. Lloyd, S.P.: Least squares quantization in pcm. IEEE Transactions on Information Theory **28**(2), 129–137 (1982). https://doi.org/10.1109/TIT.1982.1056489
9. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(Nov), 2579–2605 (2008)
10. Mahn, A., Topper, D., Quinn, S., Marron, J.: Getting started with the nist cybersecurity framework : (2021). https://doi.org/10.6028/nist.sp.1271
11. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
12. Nguyen, N., Chbeir, R., Exposito, E., Aniorte, P., Trawiński, B.: Computational collective intelligence (2019). https://doi.org/10.1007/978-3-030-28374-2
13. Probierz, B., Kozák, J., Hrabia, A.: clustering of scientific articles using natural language processing. Procedia Computer Science **207**, 3449–3458 (2022). https://doi.org/10.1016/j.procs.2022.09.403
14. Quevedo, E., Arguelles, A.P., Rodriguez, A., Yero, J., Pienta, D., Cerny, T., Rivas, P.: Creation and analysis of a natural language understanding dataset for dod cybersecurity policies (CSIAC-DoDIN v1.0) pp. 1–8 (2023). https://doi.org/10.1109/csci62032.2023.00021
15. Rose, L., Puranik, T.G., Mavris, D.N.: Natural language processing-based method for clustering and analysis of aviation safety narratives. Aerospace **7**, 143 (2020). https://doi.org/10.3390/aerospace7100143
16. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and Intelligent Laboratory Systems **2**(1-3), 37–52 (1987). https://doi.org/10.1016/0169-7439(87)80084-9