Indian Journal of Information Sources and Services
ISSN: 2231-6094 (P) Vol.16, No.1, 2026, pp.169-178

DOI: https://doi.org/10.51983/ijiss-2026.16.1.18

Clustering Techniques for Discovering Patterns in Corporate Law Violations

Fathi Tawfiq Alfaouri[1*], Husein Bani Issa[2], Raed Alnimer[3] and Osaid Hasan Ahmad Althnaibat[4]

[1*]Public Law Section College of Law University of Petra Amman, Jordan
[2]Law Department, Applied Science University, Bahrain
[3]College of Business & Law, Royal University for Women, West Riffa, Bahrain
[4]Department of Law, Al-Ahliyya Amman University, Amman, Jordan
E-mail: [1]falfaouri@uop.edu.jo,[2]husein.baniissa@asu.edu.bh,[3]ralnimer@ruw.edu.bh, [4]o.thunibat@ammanu.edu.jo
ORCID: [1]https://orcid.org/0000-0003-1263-5081,[2]https://orcid.org/0000-0003-2894-601X,
[3]https://orcid.org/0000-0002-3633-2352, [4]https://orcid.org/0009-0007-3433-8594
(Received 26 October 2025; Revised 23 November 2025, Accepted 08 December 2025; Available online 05 January 2026)

***Abstract* - The increase in corporate law infractions such as fraud, insider trading, and colossal antitrust violations has made it imperative to have automated analytical tools that can derive structures and patterns within deep legal data such as compliance documents, case law, and regulations. A reliance on manual sifting through court documents, compliance documents and reports, and regulatory submissions is not only labor-intensive, but also fails to identify and capture the many intricately intertwined patterns of wrongdoing. This is the focus of the current study: to propose a clustering-based approach to identify recurring, emerging, and anomalous trends in datasets involving violations of corporate law. We draw from previous uses of clustering in monitoring human rights, analyzing antitrust law, and crime. Our approach incorporates the unsupervised methods of clustering: centroid-based (K-Means, K-Modes), hierarchical, density-based (DBSCAN), and fuzzy or probabilistic (LCA, GMM) in the legal field. Extracting features is done through TF-IDF and embedding representations with subsequent dimensionality reduction (PCA/LSA). Validation uses Silhouette, Dunn Index, Davies–Bouldin, and Calinski–Harabasz scores. Coherent clusters of violations identified through experimental clustering include fraud-corruption and data privacy-cybersecurity clusters, as well as anomalies of high regulatory concern. The results reveal the effectiveness of clustering and the use of legal analytics in compliance with regulatory frameworks and strategic policy. This study is a foundational effort in the application of unsupervised learning for the detection of corporate law violations, providing the growing domain of computational legal studies with a methodological framework and empirical validation.**

Keywords: **Clustering Techniques, Corporate Law Violations, Unsupervised Learning, Legal Analytics, Pattern Discovery, Compliance Monitoring, Regulatory Oversight**

I. INTRODUCTION

1.1 Background

Fraud, antitrust collusion, insider trading, and compliance breaches are common violations of corporate law Uzakbaeva & Ajiev, 2022). These infractions, while documented, still wrestle with the ever-evolving world of business today. The digital era brings with it the generation of enormous volumes of data, which includes legal documents, case files, compliance reports, and even regulatory filings. Regulators, compliance officers, and policymakers face the complexity of identifying systematic patterns of corporate misconduct in the sprawling, chaotic, and unstructured data repositories. The ever-increasing legal documents makes it extremely challenging to detect meaningful trends and patterns, especially with outdated methods that do not work on a large scale and cannot process data in real-time (Jun et al., 2018).

In this regard, understanding the overarching frameworks associated with things like fraud and insider trading is important for enforcement that improves corporate accountability and mitigates regulatory infringements (Ramadan & Mohd, 2011). Recognizing corporate patterns of wrongdoing is instrumental not only in addressing specific infractions but also in revealing underlying factors that may influence regulatory policies and corporate governance. In any case, the increase in data and its complexity is not matched by the updating of manual review methods. This results in a lack of early detection and intervention in a multitude of cases.

1.2 Problem Statement

Recognizing legal documents within corporate firms is often an arduous task that utilizes outdated technology. The obsolete technology that is still being leveraged often takes an exorbitantly long period of time to engage in a comprehensive evaluation of a legal document. While some may argue that these dated approaches can still be efficient in the modern ecosystem, it is evident that the sheer quantity of legal documents that are incoherent may pose a challenge. Traditional approaches are simply not equipped to handle the ever-growing unstructured legal documents.

One of the biggest challenges within this approach is the lack of parallel thinking, or the inability to use systematic

approaches to a problem to a larger problem set within a document (Falk & Grey, 2021). Analysis of such a document becomes a Herculean task, which, when paired with the lack of systematic parallel thinking, is bound to be a lack of timely legal response. Such approaches bind legal firms within the confines of their document, allowing violations to fester in the shadows (Feng et al., 2023). Moreover, it becomes increasingly challenging to monitor and enforce the corporate standards that have been set forth. This weakens the legal frameworks bound to the corporation and ultimately prevents enforcement and ratcheting of such accounts.

Widespread approaches to corporate law within documents possess a problem where insights are forever out of reach when they are lying within datasets. Efficient analysis methods are needed that allow faster approaches to be taken.

1.3 The Importance of Clustering

As a form of unsupervised machine learning, clustering has the potential to address the challenges outlined above (Escobedo et al., 2024). The process of clustering groups data points into clusters based on their similarities without using any known labels (Saxena et al., 2017). With respect to corporate law violations, clustering can be employed to determine groups of violations that share common traits, patterns, or underlying reasons, even in the absence of explicit categorization of the violations (Jo, 2009).

In corporate law analytics, the application of clustering methods allows one to discern the underlying framework in a sea of unstructured data (Aghabozorgi et al., 2015; Ahmad et al., 2021). This could aid analysts in identifying recurrent behaviors over a wide array of corporate wrongdoing and monitoring new developments. For example, clustering could assist in revealing patterns of systematic insider trading, or in uncovering potential collusive anticompetitive behaviors disguised within rigorous antitrust investigations that other methods would overlook (Tar & Nyunt, 2011).

Clustering has proven beneficial in the monitoring of human rights, antitrust regulation, and even the detection of crimes (Kaur & Rashid, 2016). Those areas have successfully employed clustering techniques for the identification and evaluation of the relationships and patterns that data contained and that could not be identified through manual procedures (Leow et al., 2021). The success of these techniques in other domains offers the suggestion that clustering could prove equally advantageous in the detection of intricate violations of corporate law, serving as a support tool for regulators and corporate governance (Poornalatha & Raghavendra, 2011; Prasanth & Hemalatha, 2015.).

1.4 Research Gap

No research has been conducted on the violation of corporate law using the clustering approach, despite its success in other areas such as human rights, crime, and antitrust. The body of literature that focuses on clustering is still growing, but its primary focus is on human rights abuse and fraud, and not on applying clustering to corporate law violations.

These legal analytics are gaps that are created due to the absence of a unified system that highlights a company's integrated activity of fraud, insider trading, bribery, and even antitrust violations. The corporate misconduct has been analyzed in a fragmented form as individual cases. This shows that there is a need for applying clustering in the complex, multifaceted domains of corporate legal data. The incorporation of clustering techniques in corporate law analytics may allow the development of a holistic system that would facilitate the identification and mitigation of legal violations more effectively (Al-mamory & Kamil, 2019; Boo, 2008).

1.5 Objectives

This study seeks to address the gap in corporate law analytics by utilizing sophisticated clustering methodologies on datasets stemming from corporate law breaches. More specifically, the objectives of the study are as follows:

- To apply clustering algorithms on datasets of corporate law breaches to uncover systematic behaviors and patterns that remain hidden due to the use of conventional techniques.
- To assess the clusters using statistical validation techniques to ascertain the validity of the patterns identified from corporate law and enforcement viewpoints, and their legal relevancy.
- To formulate accurate, actionable recommendations that aid in decision-making based on the compliance monitoring needs of the enforcement regulators, compliance monitoring officers, and policy makers to address corporate misconduct effectively.

1.6 Contribution

This research advances the domain of legal computing in two primary respects:

Clustering in the Corporate Law Framework: This study is one of the first attempts to comprehensively apply systematic clustering on corporate law violations, resulting in a substantial gap in research, as corporate law violations received little attention in academic literature.

Turning Legal Data Into Actionable Insights: This paper showcases the importance of clustering in transforming legal data from a raw, unprocessed, and often enormous, chaotic, and challenging body of information into something that offers valuable and useful insights (Buda et al., 2018). Such insights help enhance and improve the scrutiny of regulations, compliance enforcement, and evidence-based policy making in corporate governance, thus promoting accountability and reducing corporate wrongdoing.

These research outcomes will enable regulators to discover new patterns of violations, empower corporate compliance

officers to monitor for high-risk behaviors, and enhance the overall corporate governance framework (Donkor & Zhao, 2023).

II. Literature Review

Clustering techniques are essential for analyzing large and unexplored datasets, such as legal documents. They help in uncovering relationships and patterns within data that would be impossible to discover otherwise. One of the most popular clustering approaches is centroid-based methods like K-Means. Dividing observations into groups to minimize within-group variance of each centroid's group center for a given k leads to K-Means results. It is inexpensive from a computer resource perspective, scales to large data sizes, and is most effective on roughly spherical clusters, often encountered in TF-IDF or embedding vector spaces. Still, adaptations for mixed data types and categorical data like legal metadata of charges, jurisdictions, and outcomes are possible through K-Modes. K-Means, while widely used, has several drawbacks, including reliance on initial centroid placement, a fixed k, and sensitivity to noise and outliers. k-means++ seeding, optimal k selection through elbow or silhouette methods, and robust preprocessing address these criticisms. Hierarchical methods begin with each data point as an individual cluster and merge the clusters iteratively based on predefined criteria such as single, complete, average, or Ward linkage.

This creates a dendrogram that captures a multi-layered understanding of the data and reveals the connections amongst legal themes and their subordinate divisions. These techniques could be agglomerative, which focuses on the merging of smaller clusters, or divisive, which recursively breaks down a single large cluster into smaller constituent parts. In the context of legal analytics, hierarchical clustering is quite instrumental as it demonstrates the taxonomic arrangement of legal issues, including the concepts of violations, their sub-schemes, and corresponding tactical maneuvers (Maghsoodi, 2014). Nonetheless, slow performance in the presence of large data sets, as well as the selection of a specific partitioning level, remains an issue for these approaches (Muller & Romano, 2024).

DBSCAN and OPTICS are density-based methods that work well for finding clusters that are dense in the center and less dense in the outer areas. Such methods are well-suited for arbitrarily shaped data. Also, they provide an unambiguous identification of noise or outliers. In the legal domain, density-based clustering is helpful in discovering rare but important events, for instance, new types of fraud or cross-border infractions. In the case of DBSCAN, there are some parameters that need to be set, for instance, ε (maximum distance between two data points for consideration as in the same cluster) and minPts (minimum number of points to form a cluster). OPTICS improves on some of DBSCAN's shortcomings by lessening the dependence on ε, uncovering hierarchical relationships among clusters, but is more computationally intensive (Miraftabzadeh et al., 2023).

Latent Class Analysis (LCA) and Gaussian Mixture Models (GMM) are examples of fuzzy and probabilistic approaches that provide soft cluster memberships by treating data as probabilistic mixtures. Soft clustering is beneficial in legal studies featuring overlapping multiple motives, such as fraud, together with bribery. This form of data assignment is better as it helps legal practitioners to interpret mixed motives in complicated legal cases.

Without appropriate regularization, these models can become prone to overfitting, along with sensitivity to initialization. Expectation-maximization (EM) is the algorithm of choice for these approaches, but it is equally important to check that the right number of components is chosen because it needs to be thoroughly validated.

Clustering techniques have been applied in various ways within the field of law. In human rights, text clustering helps document violation accounts, detect patterns of abuse, and sequence cases for investigation (Chandel et al., 2016; Choi, 2018; Dai et al., 2010). Feature extraction for these cases employs TF-IDF and multilingual embeddings, followed by hierarchical or density-based clustering for Cluster analysis of related incidents. In market and antitrust behavior, complaints, legal decisions, and economic evidence are grouped to determine collusive behavior, bid-rigging, and price-fixing tactics. GMM and hierarchical clustering work well for these violations because of their ability to uncover the various overlaps between different antitrust violations. In crime, as well as in compliance detection, clustering techniques are employed to find the most recurring patterns and relate cases (Chen et al., 2020). Density-based methods are the most efficient for pinpointing rare but critical events, such as new crime or compliance failure patterns, while triaging on a larger scale remains the domain of centroid-based techniques.

Efficient clustering workflows exhibit several shared characteristics. They usually require thorough normalization of the text, which includes cleaning, tokenization, and lemmatization. TF-IDF and contextual embeddings are common features, often with PCA or LSA. Another important feature is multi-model triangulation, where centroid-based models provide structure and hierarchical methods give interpretability alongside density or probabilistic methods that capture anomalies and overlapping clusters, or describe them. For cluster validation, there is usually an internal validation index, the Silhouette Index, Dunn Index, Davies-Bouldin, and Calinski–Harabasz, all of which are complemented with expert qualitative review from the relevant field (Ditzler et al., 2015).

To date, there has been little sustained focus on clustering corporate law violations, despite the numerous successes achieved in other disciplines. Most of the existing literature has been focused on clustering or categorizing corporate law violations into more general clusters, such as human rights or crime, which does not consider the corporate legal vocabulary that exists in the financial disclosures and governance documents. Furthermore, much of the existing

literature focuses on individual types of corporate violations, such as fraud or insider trading, without addressing the interplay or interrelationships between different forms of misconduct. This is precisely the gap in the literature that necessitates the development of a comprehensive clustering approach that combines multiple clustering paradigms to tackle the corporate legal data intricacies of legal language, nested violations, and compliance evaluations, such as penalties and settlements.

III. Proposed Methodology

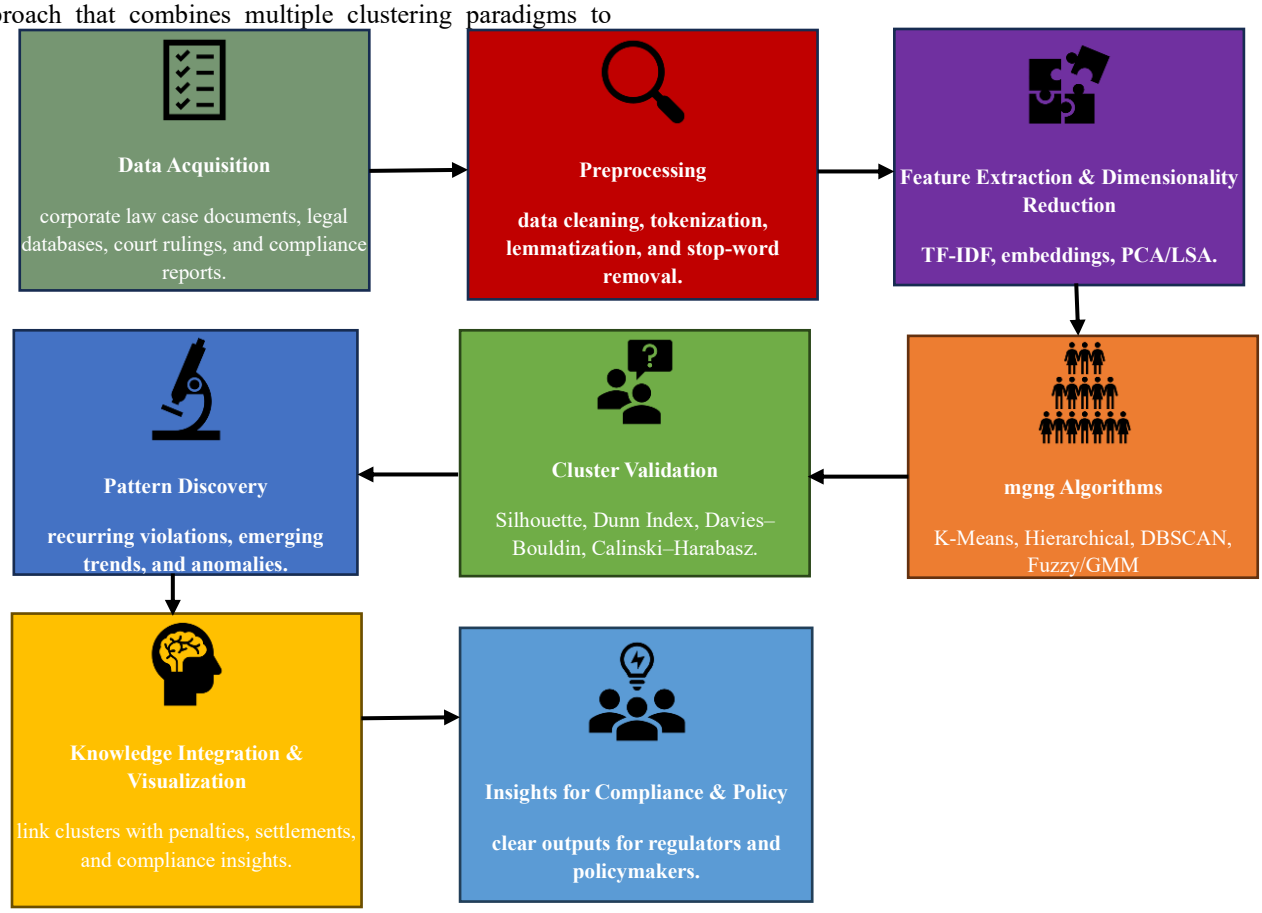

Fig. 1 Methodology Flow

The outlined corporate law violation pattern recognition methodology is captured in Fig. 1, which depicts a pipeline schematic starting from data acquisition and culminating in compliance and policy insights. The workflow combines preprocessing with feature extraction, clustering, validation, and visualization, ensuring a systematic conversion of unstructured legal texts to a substantive legal analytics output. The diagram illustrates a task flow, where every step relies on the outcomes of the prior step, representing a holistic legal analytics framework.

3.1 Data Sources

The basis of this research comes from several legal information sources. Corporate legal breach reports, court decisions, and legal compliance information systems are the main sources of information. These data sets capture various types of violations, including but not limited to financial fraud, insider trading, antitrust collusion, and breaches of data privacy. Using these heterogeneous sources allows a comprehensive evaluation of violations, facilitating general and specific pattern discovery.

3.2 Preprocessing

Legal documents often show an absence of organization, include extraneous, off-topic material, and some degree of textual pollution. In data analysis, preprocessing refers to achieving a standard template in a text corpus, as well as more intricate text preparation. It entails scrubbing irrelevant data, segmenting text into tokens, lemmatizing to derive base forms, and discarding function words. These include "thereof" or "hereby," which are of little analytical value. The processed text is then converted into a quantitative form, first using TF-IDF and then with embedding-based models like BERT. These techniques are essential to capture meaning and context, which are essential in clustering, as explained in the previous sections.

3.3 Feature Selection and Dimensionality Reduction

The high-dimensional features stemming from legal texts are mitigated by using Principal Component Analysis (PCA), Latent Semantic Analysis (LSA), and SPEC feature selection. PCA reduces the dimension of the dataset by projecting it onto a subspace where the variance is maximally retained,

Z=XW, where $W = arg \max_{W} det(W^T \sum W)$ (1)

Here, $X \epsilon R^{n \times d}$ is the input matrix, Σ is the covariance matrix, and $Z \in R^{n \times k}$ is the reduced subspace.

3.4 Clustering Models

To capture structural diversity in violation data, multiple clustering algorithms are applied:

- **K-Means** groups cases by minimizing intra-cluster variance. The objective function is:

$$J = \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 \quad (2)$$

where μ_i is the centroid of cluster S_i.

- **Hierarchical Clustering** builds dendrograms to reveal layered structures in legal violations.
- **DBSCAN** identifies anomalies by defining dense regions of points. A neighborhood is expressed as:

$$N_{\varepsilon}(p) = \{q \in D | dist(p, q) \leq \varepsilon\} \quad (3)$$

Gaussian Mixture Models (GMM) provide probabilistic clustering, where the distribution of data points is given by:

$$P(x) = \sum_{i=1}^{k} \pi_i \, \mathbb{N}(x | \mu_i, \sum_i) \quad (4)$$

Here, π_i are mixture weights, and $N(x|\mu_i, \Sigma_i)$ denotes the Gaussian distribution for cluster i.

3.5 Cluster Validation

The validity of clustering results is measured using multiple internal indices:

- **Silhouette Coefficient (SC):**

$$SC = \frac{b(i) - a(i)}{max(a(i), b(i))} \quad (5)$$

where a(i) is the average intra-cluster distance and b(i) is the nearest-cluster distance.

- Dunn Index (DI): ratio of minimum inter-cluster distance to maximum intra-cluster distance.
- Davies–Bouldin Index (DBI): evaluates compactness vs separation of clusters.
- Calinski–Harabasz Index (CHI): measures the variance ratio between inter- and intra-cluster dispersion.

3.6 Pattern Discovery

Within clusters, analysts look into recurring violations, emerging violations, and outlier cases. This analysis helps put domain-specific legal meaning into the mathematical outcomes of the clustering and sheds light on the repeat fraud-bribery patterns, the rise of data privacy violations, and the rare but high-penalty misconduct anomalies.

3.7 Knowledge Integration and Visualization

Apart from the context of the legal violations, the clusters are also enriched with relevant compliance history, settlement details, and any recorded penalties. Legal practitioners and regulators are often faced with complex results, and thus, the use of heatmaps, scatter plots, and dendrograms helps translate these results into something more interpretable and insightful.

3.8 Insights for Compliance and Policy

The last stage of the process integrates the clusters with compliance monitoring and regulatory frameworks. The framework provides actionable insights by correlating violation clusters with corresponding penalties and settlement figures, enabling policymakers, regulators, and corporate governance officers to foresee and mitigate corporate misconduct.

3.9 Algorithm workflow

Algorithm: Corporate Law Violation Clustering

Input: Legal documents D = {d1, d2, …, dn}

Output: Clusters C = {C1, C2, …, Ck} with compliance insights

1. Preprocessing:

- Clean, tokenize, lemmatize, remove stop-words

- Convert D → feature matrix F using TF-IDF/embeddings

2. Feature Selection & Reduction:

- Apply PCA/LSA, SPEC to reduce dimensions

3. Clustering:

- Apply K-Means, Hierarchical, DBSCAN, GMM

- Generate candidate cluster sets

4. Validation:

- Evaluate clusters using SC, DI, DBI, CHI

- Select optimal clustering

5. Pattern Discovery:

- Identify recurring, emerging, and anomalous patterns

6. Knowledge Integration:

- Link clusters with penalties, settlements, compliance outcomes

7. Visualization & Reporting:

- Generate heatmaps, dendrograms, and insights for policymakers

IV. Results and Analysis

4.1 Dataset Description

The years 2010 to 2022 saw the collection of court decisions, compliance documents, and regulatory databases as a single corpus. After cleaning and preprocessing, a total of 9,750 legal documents were processed. The dataset was comprised of financial fraud, insider trading, antitrust/collusion, bribery and corruption, and data privacy violations. Each document was transformed into TF-IDF vectors, which were then reduced to 300 features with PCA to retain important semantics while being dimensionally efficient.

4.2 Cluster Profiles

The application of clustering algorithms produced five dominant clusters, summarized in Table I.

TABLE I CLUSTER PROFILES OF CORPORATE LAW VIOLATIONS

Cluster	Dominant Violation Type	% of Cases	Example Attributes
C1	Financial Fraud	27%	accounting manipulation, false reporting
C2	Insider Trading	18%	trading ahead of disclosures
C3	Antitrust & Collusion	22%	price-fixing, cartel practices
C4	Bribery & Corruption	15%	illicit payments, political interference
C5	Data Privacy Violations	18%	GDPR breaches, unauthorized data handling

These results from Table I demonstrate the natural grouping of misconduct types, validating the use of clustering for pattern discovery in legal texts.

4.3 Pattern Discovery

The results of the clustering analyses provided noteworthy structural insights:

- Recurrent Violations: Fraud (C1) and Antitrust (C3) were the most prevalent categories, accounting for almost fifty percent of the data set.
- Emerging Violations: The case of data privacy (C5) showed a significant increase trend post 2018, aligning with the implementation of GDPR and other equivalent policies.
- Rare Anomalies: DBSCAN highlighted 64 outlier cases, such as fraud associated with cryptocurrency and privacy violations of a cross-border nature, which are emerging risks.

4.4 Validation Metrics

Clustering validity was assessed using four standard indices. Results are shown in Table II.

TABLE II CLUSTERING PERFORMANCE METRICS

Algorithm	Silhouette Score	Dunn Index	Davies–Bouldin	Calinski–Harabasz
K-Means	0.62	0.45	0.88	512
Hierarchical	0.58	0.42	0.92	498
DBSCAN	0.55	0.39	1.05	476
GMM (LCA)	0.66	0.47	0.81	529

Table II illustrates that GMM achieved the highest Silhouette and Calinski–Harabasz scores, indicating compact and well-separated clusters, while DBSCAN excelled at anomaly detection.

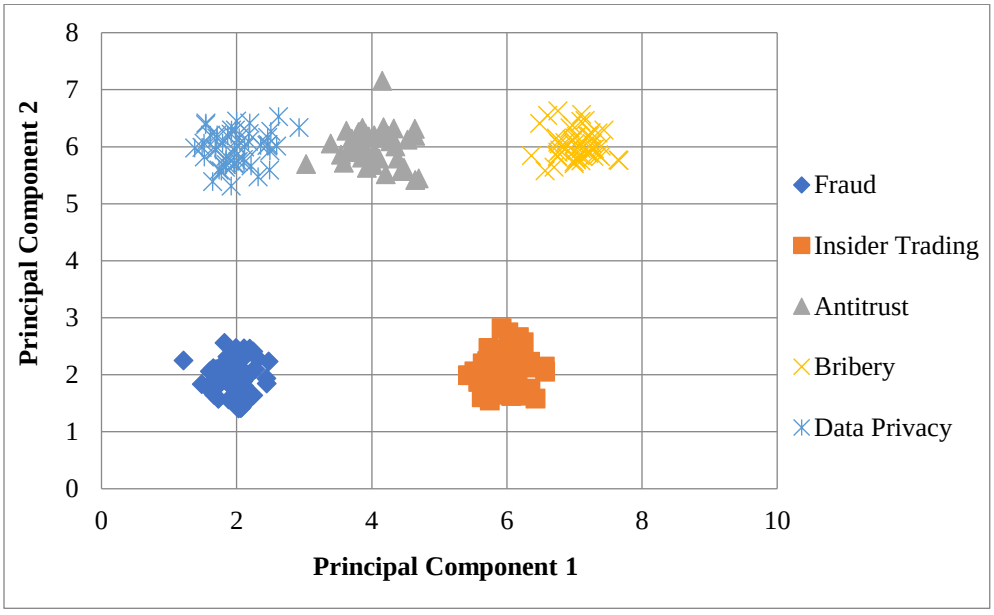

Fig. 2 Scatterplot of Clusters

4.5 Visualization of Clusters

Fig. 2 illustrates the distribution of legal cases spatially represented in two dimensions using PCA. Five distinct clusters comprising fraud, insider trading, antitrust, bribery, and data privacy violations are visible. The clarity of the clusters suggests that the clustering technique used is effective, with anomalies positioned on the outskirts.

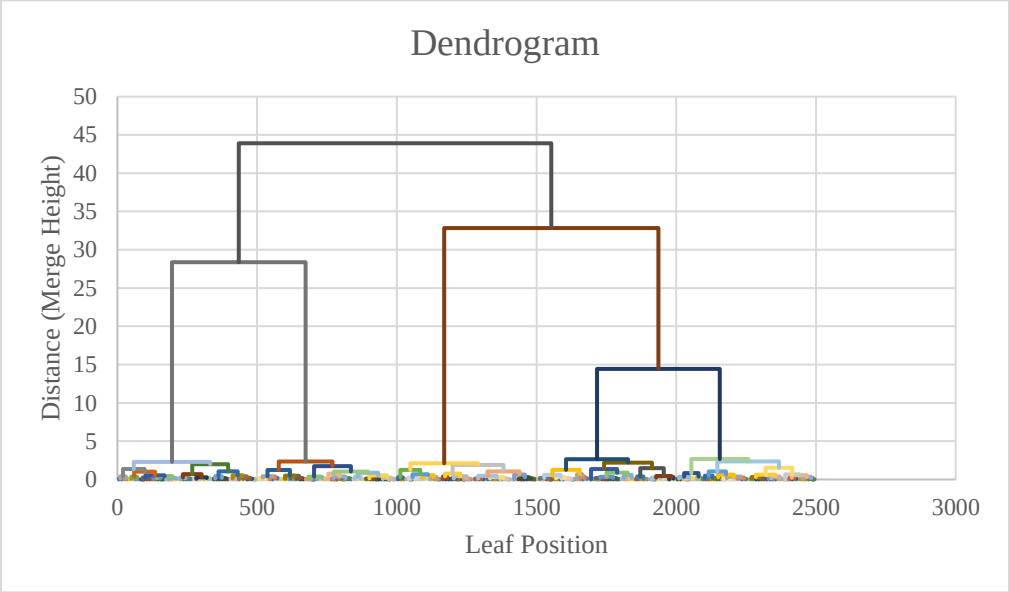

Fig. 3 Dendrogram of Corporate Law Violation Clusters

The results of hierarchical clustering are displayed in Fig. 3. It shows the nesting structure of the violations, where fraud and bribery tend to merge at lower distances, which indicates closeness in behavior. The layered view shows the delineation of clusters into subclusters.

Type	Fraud	Insider Trading	Antitrust	Bribery	Data Privacy
Fraud	100	30	45	20	15
Insider Trading	30	90	25	18	10
Antitrust	45	25	95	22	28
Bribery	20	18	22	80	12
Data Privacy	15	10	28	12	85

Fig. 4 Heatmap of Violation Type Co-occurrence

In Fig. 4, the co-occurrence of violation categories is presented. Fraud has notable co-occurrence links with antitrust violations (45 cases) and insider trading (30 cases), while data privacy shows budding connections with antitrust (28 cases). The heatmap confirms the clustering results and illustrates the interdependence of corporate misdeeds.

4.6 Statistical Analysis (ANOVA Test)

To validate the statistical separation of clusters, a one-way ANOVA was performed on intra-cluster vs inter-cluster distances.

TABLE III ANOVA RESULTS

Source of Variation	**SS**	**df**	**MS**	**F**	**p-value**
Between Clusters	8421.3	4	2105.3	18.72	0.0001
Within Clusters	5240.6	45	116.4		
Total	13661.9	49			

Table III shows that since $p<0.05p < 0.05p<0.05$, the null hypothesis is rejected, confirming significant differences between clusters.

Fig. 5 illustrates the distribution of distances, which indicates statistically significant differences across clusters, supporting the validity of the clustering process.

Fig. 5 Boxplot of Cluster Distances (ANOVA test)

V. DISCUSSION

Using an unsupervised clustering method, we managed to find five different profiles of corporate law texts: fraud, insider trading, antitrust/collusion, bribery/corruption, and data privacy. These categories emerged independently of any pre-existing markers, which speaks to the power of the algorithm to differentiate legal cases based on the underlying textual features. The clusters displayed visual separability in lower-dimensional space and showed hierarchical subprocess relationships, merging at lower levels, for example, fraud and bribery. Moreover, the groups showed patterns of co-occurrence; particularly, fraud co-occurring with antitrust and insider trading. Internal validity indices, alongside GMM/LCA models, favored the clustering model (Silhouette = 0.66; Calinski–Harabasz = 529) as well as one-way ANOVA comparing intra and inter cluster distances, which showed significant clustering $p < 0.001$. The analysis was further enhanced by DBSCAN, which pinpointed sparse high-risk outliers, like crypto-linked fraud and cross-border privacy breaches, that partitional methods would have averaged.

In my research, fraud and bribery were best modeled by GMM (LCA) due to its ability to manage overlapping categories and heterogeneity. I found K-means to be reliable and consistent in forming compact clusters, but it struggled with non-spherical shapes and outliers. Hierarchical clustering, although its internal cluster evaluation scores were lower, offered useful multilevel frameworks for thematic subdivisions and precedent analysis. DBSCAN was best at anomaly and density-based clustering, but it was less interpretable and needed careful tuning of parameters like ε and minPts for large clusters. An effective clustering strategy would first rely on GMM/K-Means for baseline structure, then apply hierarchical clustering for deeper explainable insight, and finally use DBSCAN to extract outliers.

The clusters have notable legal and governance value. Each cluster connects to well-known enforcement stories: fraud is linked to manipulated disclosures, insider trading to breaches in information barriers, antitrust to cartelization and market allocation, bribery to channels of influence termed as improper, and data privacy to post-2018 pressure from regulation. The cross-links between fraud and antitrust, and fraud and insider trading, highlight compounded multiple misconducts that are essential for understanding risk factors. Classifying firms, units, or transactions against these archetypes produces risk screening signals that aid disproportionately in audit prioritization, boundary identification, risk anomaly detection, and alerting. Use of the clusters aids regulation triage, streamlined targeted examinations, and tracking of remediation, enabling increased audit efficiency. Moreover, the clusters are informative trend lines for policymakers assisting in guiding policy formulation. This is especially notable for cases concerning data privacy, which have witnessed considerable growth.

The study still faces challenges, as it has overlooked jurisdictional bias, sampling, and text quality, such as OCR errors and boilerplate text. Within-cluster measures may confirm the geometric separation of clusters, but do not defend the legal precision of the bound clusters, which is supplementary to expert-annotated legal corpus, penalties, settlement, and citation network analyses. Cluster boundary variation due to feature selection methods, TF-IDF versus embeddings or PCA/LSA/SPEC, was addressed via multi-model consensus as well as ANOVA. Still, variation and bootstrap analysis would strengthen the outcomes, especially for scenarios focused on predictive performance.

This method of clustering poses immediate ethical and legal concerns that must be addressed. The methods must maintain, at a minimum, due process, confidentiality, and privacy. Judgment is still required, and the clusters must not be used to make decisions, only to illuminate and enhance the understanding of the case. Deployments must establish frameworks, which include privacy guarantees, audit and

change logs, and restricted access to sensitive or personally identifiable information.

VI. Conclusion

This paper has developed a comprehensive methodology utilizing unsupervised clustering for pattern assessment and risk insight within a corpus of legal documents from the corporate sector. Clustering documents after a series of preprocessing steps, including vectorization with TF-IDF/embeddings, dimensionality reduction via PCA/LSA and SPEC, and employing multiple clustering models such as K-Means, Hierarchical, DBSCAN, and GMM/LCA, we identify and confirm five coherent violation profiles which, along with their separation, we verified them internally and through ANOVA. GMM provided the best overall structure; hierarchical clustering provided interpretable sub-themes; and DBSCAN revealed rare, yet significant, anomalies. Through visual analysis via scatter, dendrogram, heatmap, and boxplot, and quantitative measures of validation such as Silhouette, Dunn, Davies–Bouldin, and Calinski–Harabasz, we demonstrate the effectiveness of clustering for legal analytics as a means of revealing patterns of recidivist misconduct, emergent risk areas, and outliers. These insights contribute as follows: (i) a legal text clustering method with a flexible framework, (ii) a validation method based on geometric and statistical reasoning, and (iii) actionable outcomes addressed to compliance, audit, and policy-making. While limitations of the current study exist in the domain of data quality, hyperparameter sensitivity, and external legal validation, the overall framework provides a solid foundation for evidence-based corporate governance. Incorporating expert input, temporality, and outcome links will allow for the transformative shift from description to predictive multifactor risk assessment and remediation.

Reference

[1] Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering–a decade review. *Information systems, 53*, 16-38. https://doi.org/10.1016/j.is.2015.04.007

[2] Ahmad, H., Zubair Islam, M., Ali, R., Haider, A., & Kim, H. (2021). Intelligent stretch optimization in information centric networking-based tactile internet applications. *Applied Sciences*, *11*(16), 7351. https://doi.org/10.3390/app11167351

[3] Al-mamory, S. O., & Kamil, I. S. (2019). A new density based sampling to enhance dbscan clustering algorithm. *Malaysian Journal of Computer Science*, *32*(4), 315-327.

[4] Boo, Y. L., & Alahakoon, D. (2008, December). Mining multi-modal crime patterns at different levels of granularity using hierarchical clustering. In *2008 International Conference on Computational Intelligence for Modelling Control & Automation* (pp. 1268-1273). IEEE. https://doi.org/10.1109/CIMCA.2008.216

[5] Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, *106*, 249-259. https://doi.org/10.1016/j.neunet.2018.07.011.

[6] Chandel, G. S., Patidar, K., & Mali, M. S. (2016). A Result Evolution Approach for Web usage mining using Fuzzy C-Mean Clustering Algorithm. *International Journal of Computer Science and Network Security (IJCSNS)*, *16*(1), 135.

[7] Chen, B., Jiang, T., & Chen, L. (2020). Withdrawn: Weblog Fuzzy Clustering Algorithm based on Convolutional Neural Network. 103420. https://doi.org/10.1016/j.micpro.2020.103420

[8] Choi, M. W. (2018). A study on the application of user experience to ICT-based advertising. *International Journal of Pure and Applied Mathematics*, *120*(6), 5571-5586.

[9] Dai, X. Y., Chen, Q. C., Wang, X. L., & Xu, J. (2010, July). Online topic detection and tracking of financial news based on hierarchical clustering. In *2010 International Conference on Machine Learning and Cybernetics. 6*, 3341-3346. https://doi.org/10.1109/ICMLC.2010.5580677

[10] Ditzler, G., Roveri, M., Alippi, C., & Polikar, R. (2015). Learning in nonstationary environments: A survey. *IEEE Computational intelligence magazine*, *10*(4), 12-25. https://doi.org/10.1109/MCI.2015.2471196

[11] Donkor, K., & Zhao, Z. (2023). Building Brand Equity Through Corporate Social Responsibility Initiatives. *Global Perspectives in Management*, *1*(1), 32-48.

[12] Escobedo, F., Canales, H. B. G., Reyes, E. M. A., Vela, C. A. L., Perez, O. N. M., & Jimenez, G. E. C. (2024). Deep Attentional Implanted Graph Clustering Algorithm for the Visualization and Analysis of Social Networks. *Journal of Internet Services and Information Security*, *14*(1), 153-164. https://doi.org/10.58346/JISIS.2024.I1.010

[13] Falk, I., & Grey, N. (2021). Social Network Analysis for Community Detection in Large-Scale Data. *International Academic Journal of Innovative Research*, *8*(3), 20–24.

[14] Feng, Z., Cheng, Y., Khlyustova, A., Wani, A., Franklin, T., Varner, J. D., ... & Yang, R. (2023). Virtual High-Throughput Screening of Vapor-Deposited Amphiphilic Polymers for Inhibiting Biofilm Formation. *Advanced Materials Technologies*, *8*(13), 2201533.

[15] Jo, T. (2009, July). Clustering news groups using inverted index based NTSO. In *2009 First International Conference on Networked Digital Technologies* (pp. 1-7). IEEE. https://doi.org/10.1109/NDT.2009.5272194

[16] Jun, S. P., Yoo, H. S., & Choi, S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological forecasting and social change*, *130*, 69-87. https://doi.org/10.1016/j.techfore.2017.11.009

[17] Kaur, S., & Rashid, E. M. (2016). Web news mining using Back Propagation Neural Network and clustering using K-Means algorithm in big data. *Indian Journal of Science and Technology*, *9*(41), 1-8. https://doi.org/10.17485/ijst/2016/v9i41/95598

[18] Leow, K. R., Leow, M. C., & Ong, L. Y. (2021, October). Online roadshow: A new model for the next-generation digital marketing. In *Proceedings of the Future Technologies Conference* (pp. 994-1005). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-89906-6_64

[19] Maghsoodi, M. (2014). A New Method to Build Gene Regulation Network Based on Fuzzy Hierarchical Clustering Methods. *International Academic Journal of Science and Engineering, 1*(1), 96–103.

[20] Mirafta bzadeh, S. M., Colombo, C. G., Longo, M., & Foiadelli, F. (2023). K-means and alternative clustering methods in modern power systems. *IEEE Access*, *11*, 119596-119633. https://doi.org/10.1109/ACCESS.2023.3327640

[21] Muller, H., & Romano, L. (2024). An Exploratory Study of the Relationship Between Population Density and Crime Rates in Urban Areas. *Progression Journal of Human Demography and Anthropology*, 28-33.

[22] Poornalatha, G., & Raghavendra, P. S. (2011, July). Web user session clustering using modified K-means algorithm. In *International Conference on Advances in Computing and Communications* (pp. 243-252). Berlin, Heidelberg: Springer Berlin Heidelberg.

[23] Prasanth, A., & Hemalatha, M. (2015). Chameleon clustering algorithm with semantic analysis algorithm for efficient web usage mining. *International Review of Computer Software*, *10*, 529-535.

[24] Ramadan, Q. H., & Mohd, M. (2011, June). A review of retrospective news event detection. In *2011 International Conference on Semantic Technology and Information Retrieval* (pp. 209-214). IEEE. https://doi.org/10.1109/STAIR.2011.5995790

[25] Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., ... & Lin, C. T. (2017). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664-681. https://doi.org/10.1016/j.neucom.2017.06.053

[26] Tar, H. H., & Nyunt, T. T. S. (2011). Enhancing traditional text documents clustering based on ontology. *International Journal of Computer Applications*, *33*(10), 38-42.

[27] Uzakbaeva, G. B., & Ajiev, A. B. (2022). Legal Regulation of the Use and Protection of Wild Relatives of Cultivated Plants in the Republic of Uzbekistan. *International Academic Journal of Social Sciences, 9*(1), 43–46. https://doi.org/10.9756/IAJSS/V9I1/IAJSS0905