Journal of COMPUTER AND FORENSIC SCIENCES

CFS 2025, Vol. 4, Issue 2, pp. 30–42
https://doi.org/10.5937/jcfs4-64411
Original research paper

Comparative Analysis of Clustering Textual and Numerical Data Using the K-Means Algorithm

Sanja Raičević[1*]

[1] Ministry of Interior of the Republic of Serbia

*Corresponding author: sanjaraicevic09@gmail.com

Received: January 27, 2026 • Accepted: April 25, 2026 • Published: May 14, 2026

Abstract: This paper presents a comparative analysis of the application of the K-Means clustering algorithm on two different types of data – textual and numerical. The aim of the research was to examine the reliability, stability, and interpretability of the results when the same algorithm is applied to semantically diverse datasets. The textual data were taken from the articles of the Criminal Code of the Republic of Serbia, where clustering was performed after preprocessing and TF-IDF vectorization. The numerical data refer to traffic accident statistics from 2015 to 2021, analyzing parameters such as the number of property-damage-only accidents, the number of injured persons, and the number of fatalities.

The results showed that clustering on textual data produced a relatively clear separation of thematic groups of articles, but with a moderate silhouette coefficient value due to a high degree of semantic similarity among documents. On the other hand, clustering on numerical data demonstrated a more stable structure, where the optimal number of clusters was two, indicating the possibility of distinguishing periods with different intensity and severity of traffic accidents.

It was concluded that the K-Means algorithm provides more reliable and interpretable results for numerical data, while in the case of textual data, it requires more precise vector space modeling and possibly the application of semantic models such as Word2Vec or BERT. The paper serves as a basis for further research in the field of integrating machine learning techniques for analyzing heterogeneous data sources.

Keywords: clustering, K-Means, textual data, numerical data, TF-IDF, PCA, silhouette analysis.

1. INTRODUCTION

The modern era is characterized by an exponential growth in the volume of data generated daily across various domains – from textual documents and administrative records to numerical statistical reports and sensor data. In such an environment, the ability to efficiently discover structures and patterns within data has become one of the key tasks of modern analytics and machine learning. One of the fundamental approaches in this context is *clustering*, which refers to the grouping of data into clusters based on their internal similarity.

Among numerous clustering algorithms, K-Means occupies a central position due to its simplicity, speed, and broad applicability. It enables the identification of groups of objects that share similar characteristics, without prior knowledge of the underlying data structures. However, the effectiveness of the K-Means algorithm largely depends on the nature

Citation: S, Raičević "Comparative Analysis of Clustering Textual and Numerical Data Using the K-Means Algorithm" *Journal of Computer and Forensic Sciences*, vol. 4, no. 2, pp. 30–42, 2025, https://doi.org/10.5937/jcfs4-64411. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

of the processed data, their distribution, and the applied preprocessing methods. For this reason, it is important to examine how this algorithm performs when applied to different types of data.

The subject of this research is a comparative analysis of clustering results for textual and numerical data using the same algorithm – K-Means. The textual data were taken from the corpus of articles of the Criminal Code of the Republic of Serbia, while the numerical data were based on official traffic accident statistics from 2015 to 2021. This combination of two semantically distinct datasets allows for a deeper analysis of the K-Means algorithm's ability to recognize internal structures within data of different natures.

The aim of the paper is to determine the extent to which the K-Means algorithm produces accurate and interpretable results in both cases, as well as to highlight its limitations and potential applications in various analytical contexts. The paper presents the steps of data preparation, clustering implementation, visualization of results, and interpretation of the obtained clusters.

2. RELATED WORK

Research on data clustering has a long history and represents one of the key fields in data analysis and machine learning. The foundations of this approach were established by MacQueen [1], who defined the K-Means algorithm as a method for grouping multivariate data based on a measure of similarity.

A review of existing studies shows that clustering is a dynamic field that unifies classical statistical methods with modern deep learning approaches. In this context, the present paper builds upon previous studies that applied K-Means and PCA techniques to the analysis of traffic accident data [2], with the possibility of future expansion through the application of advanced vector models and hybrid clustering methods.

The classification and analysis of textual data pose a particular challenge within this domain, as the text must be transformed into a numerical form suitable for further processing. Havrlant and Kreinovich [3] provided a formal explanation of the TF-IDF heuristic, one of the most commonly used approaches for converting textual content into vector representations, while Wu and Xu [4] emphasized the importance of various machine learning techniques for text classification. The TF-IDF method often represents the initial step in the process of clustering textual documents, which was also the case in this study.

Verification and evaluation of the quality of the formed clusters represent an essential aspect of any research in this area. Rousseeuw [5] proposed the Silhouette Score metric as both a visual and quantitative means for assessing the trade-off between cluster compactness and separation, which later became a standard criterion for evaluating clustering results.

The application of Principal Component Analysis (PCA) has a long tradition in data analysis and dimensionality reduction. Recent studies on PCA have demonstrated its effectiveness as a natural method for visual and statistical data exploration, such as the work by Gewers et al. [6], while Dunteman [7] and Jolliffe [8] systematized the mathematical foundations of the PCA approach. In the context of these works, PCA was applied for graphical

representation of relationships among parameters affecting the number and severity of traffic accidents, thereby enabling a better understanding of variability within the dataset.

Manning, Raghavan, and Schütze [9], in their seminal work *Introduction to Information Retrieval*, defined the theoretical and practical aspects of information processing, text classification, and retrieval, establishing the framework within which modern machine learning algorithms and text analysis methods have evolved. Subsequent research inspired by MacQueen, such as the work of Xu and Tian [10], expanded this concept through a comprehensive overview of contemporary clustering methods, highlighting their advantages and limitations depending on data structure and dimensionality.

The processing of textual data has been further improved through the application of word vectorization methods based on deep learning. Mikolov et al. [11] developed the Word2Vec model, which enables the representation of words in vector space while preserving semantic proximity, whereas Pennington, Socher, and Manning [12] proposed the GloVe model as a more advanced variant that combines local and global statistical information. These models now form the foundation of modern natural language processing techniques and have the potential to significantly enhance clustering results compared to the traditional TF-IDF approach.

3. THEORETICAL FRAMEWORK AND RESEARCH METHODOLOGY

3.1. Theoretical Framework

Clustering methods represent one of the fundamental approaches within unsupervised machine learning, with the goal of identifying natural groups of objects based on their similarities or differences. Unlike supervised learning methods, where categories are predefined, clustering operates without prior knowledge of data structures, making it particularly useful for uncovering hidden patterns.

One of the most commonly used clustering algorithms is K-Means, proposed by James MacQueen (1967) [1]. This algorithm seeks to minimize the total distance of data points from their respective cluster centroids, thereby producing groups that are as homogeneous as possible. The basic principle of operation is based on defining the number of clusters K, selecting initial centroids, assigning each object to the nearest centroid, and iteratively updating centroid positions until a stable cluster structure is achieved.

The K-Means algorithm can be applied to different types of data – numerical data, where values have directly measurable relationships, and textual data, where prior vectorization is necessary to represent textual content in numerical form. For textual data, one of the most important techniques is TF-IDF (Term Frequency – Inverse Document Frequency), which quantifies the significance of words relative to their occurrence in the entire corpus. The resulting values form vectors that serve as input data for the clustering algorithm.

To reduce data complexity and improve visualization, the Principal Component Analysis (PCA) technique is often applied. PCA transforms the original set of correlated variables into a smaller number of uncorrelated principal components, retaining most of the total data variance. This approach enables better interpretation of results and clearer visualization in two- or three-dimensional space [2].

The application of clustering to the analysis of textual and numerical data allows researchers to gain deeper insights into data structures and underlying patterns. While clusters in textual datasets are formed based on semantic similarity, in numerical data they are created based on quantitative and statistical relationships among variables. This research aims to demonstrate the differences in processing methods, result quality, and cluster interpretability between these two data types using the K-Means algorithm [3].

3.2. Research Methodology

The methodological approach in this paper is based on the experimental application of the K-Means algorithm to two datasets of different natures:

1) Textual data – articles of the Criminal Code in *.txt* format,

2) Numerical data – datasets containing numerical values related to the occurrence and consequences of traffic accidents in *.csv* format.

3.2.1. Data Preparation and Preprocessing

For the textual corpus, several preprocessing stages were carried out:

– text cleaning (removal of punctuation, numbers, and stop words),

– tokenization and transformation of words into numerical vectors using TF-IDF vectorization,

– dimensionality reduction using PCA to improve interpretability and optimize computation,

data normalization to ensure that all values are within the same scale [4].

For the numerical data, standard preparation phases were conducted, including:

– identification and removal of missing values,

– data scaling and normalization,

– selection of relevant variables representing the intensity, frequency, and consequences of the observed phenomena.

3.2.2. Application of the Algorithm and Determination of the Optimal Number of Clusters

The K-Means algorithm was applied to both datasets, with systematic testing of different values of the parameter *K*. To determine the optimal number of clusters, two standard methods were used:

– The **Elbow method**, which observes the relationship between inertia and the number of clusters,

The **Silhouette Score**, which measures the quality of clustering by considering both the distance between clustered objects and the compactness within each group [5].

3.2.3. Visualization and Analysis of Results

The clustering results were visualized in a two-dimensional space using PCA, which enabled a clearer representation of the boundaries between the formed clusters. In the case of textual data, the visualization reveals semantic groupings of legal articles, while the numerical data display trends in the frequency and severity of occurrences across different time intervals [6].

3.2.4. Objective and Expected Contribution

The main objective of the methodology is to determine the extent to which the type of data (textual or numerical) influences the clustering results, as well as to evaluate the effectiveness of the K-Means algorithm in different contexts. The obtained results may contribute to the development of systematic approaches for data analysis in legal, social, and technical domains, where it is necessary to identify natural groupings and patterns without predefined labels.

4. ANALYSIS OF RESULTS

The research was conducted on two datasets of different nature – textual and numerical types. The goal was to apply the same clustering algorithm (K-Means) and compare the obtained results in order to observe differences in the way this algorithm groups data of different structures.

The textual data consisted of segments of articles from the Criminal Code of the Republic of Serbia. Each article was treated as a separate document. Before clustering, a preprocessing procedure was carried out, which included the following steps:

– removal of punctuation marks and stop words,

– conversion of text to lowercase for consistency,

– lemmatization to reduce words to their base form,

– transformation into a numerical format using TF-IDF vectorization.

The resulting TF-IDF weight matrix for the textual data served as the basis for applying the K-Means algorithm. The value of the K parameter was selected experimentally, based on the Elbow method, which determines the number of clusters at which the internal inertia stabilizes. After that, PCA analysis was performed to reduce dimensionality to two components, enabling visualization of the results. The results showed that the legal articles were grouped according to thematic similarities, meaning that sections of the law related to similar areas (e.g., criminal acts against life and body, property-related crimes, etc.) appeared within the same clusters. This confirms the applicability of the K-Means algorithm for semantic clustering of textual content.

The second part of the research was based on the analysis of numerical data related to traffic accidents. The dataset contained information on the number of accidents, fatalities, and injuries per year, as well as other relevant parameters.

Before applying the K-Means algorithm, data standardization was performed using the StandardScaler technique to ensure that all attributes had an equal impact on cluster formation. Then, the data were transformed into a two-dimensional space using PCA analysis, which allowed for visual representation of the clusters and easier interpretation of the results.

The clustering results showed that the data were grouped according to years and the severity of traffic accident consequences. It was observed that more recent years exhibited higher concentrations within clusters characterized by greater accident intensity and a higher number of severe outcomes, which may indicate changes in traffic intensity, infrastructure, or participants' behavior.

On the PCA diagram, the first component (PCA1) represents the main direction of variance in the data and includes parameters with the greatest influence on the number and severity of accidents, while the second component (PCA2) represents the remaining variance related to secondary characteristics such as the number of injured persons or local conditions. The visual analysis clearly shows a separation between periods with fewer and more accidents, allowing for a better understanding of temporal trends. [7]

4.1. Comparative Analysis of Clustering Results for Textual and Numerical Data

By comparing the clustering results of textual data from the Criminal Code and numerical data on traffic accidents, both common methodological features and differences arising from the nature of the data can be observed.

In both cases, the K-Means algorithm was applied, along with the Elbow method and Silhouette Score, to determine the optimal number of clusters. For the textual data, the optimal number of clusters was three, whereas for the numerical data, the best results were achieved with two clusters. This difference arises from the complexity of textual documents, which contain multiple semantic nuances, compared to the simpler and more directly measurable structure of numerical data.

From an interpretative perspective, clustering the textual data enabled grouping legal articles according to thematic similarity, allowing for the automatic identification of areas related, for example, to criminal acts against life, property, or public order. Clustering the numerical data, on the other hand, resulted in the identification of periods with higher or lower frequency of traffic accidents, which has direct applications in monitoring safety trends and planning preventive measures.

Both analyses highlighted the importance of applying PCA for visualizing results. Dimensionality reduction allowed for clear graphical representation of clusters in 2D space and improved understanding of the relationships between the observed objects. While clusters in textual data were more dispersed and scattered, numerical data formed more compact and precisely defined groups [8].

It can be concluded that clustering represents a universal analysis technique, applicable to both unstructured (textual) and structured (numerical) data. However, the success of the analysis largely depends on preprocessing, choice of metrics, and understanding of the data context. By combining these approaches, it is possible to build integrated systems that

simultaneously analyze legal regulations and statistical data, thereby enabling a deeper understanding of the relationship between the legislative framework and the social reality described by the data.

4.2. Quantitative Analysis and Model Stability

The results presented in Table 1 show that the K-Means algorithm demonstrates greater stability with numerical data than with textual data. For textual data, there is a higher sensitivity to the choice of parameters (number of clusters, number of components in PCA, method of vectorization), which leads to fluctuating results. In contrast, with numerical data, the algorithm consistently identifies stable boundaries between groups, allowing for more reliable interpretation.

Table 1. Quantitative analysis of clustering results.

Data Type	Optimal Number of Clusters (K)	Silhouette Score	Visual Separation	Model S tability
Textual Data	Difficult to determine (gradual decline)	0.02 – 0.045	Weak, clusters overlap	Low
Numeri-cal Data	K = 2	≈ 0.557	Clearly separated clusters	Medium–High

The main difference lies in the nature of the data itself. Numerical attributes are quantitatively defined and suitable for the distance metrics used by K-Means, whereas textual data represent complex semantic structures that cannot be easily projected into Euclidean space without loss of meaning. Even after TF-IDF transformation and PCA reduction, textual data retain semantic ambiguity. Similarity between legal articles often depends on context rather than just word frequency, leading to less clear boundaries between classes [9]. On the other hand, numerical data (such as the number of accidents, injuries, and fatalities) have a clear statistical structure, allowing the algorithm to effectively detect natural groups and identify patterns, such as differences between days with increased and decreased numbers of traffic incidents.

Table 2 presents the clustering results for the two datasets—textual and numerical—using the K-Means algorithm. The display is divided into three segments for each data type:

1) Visualization of clusters after dimensionality reduction using PCA (Principal Component Analysis),

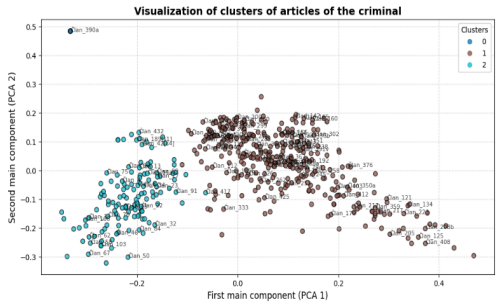
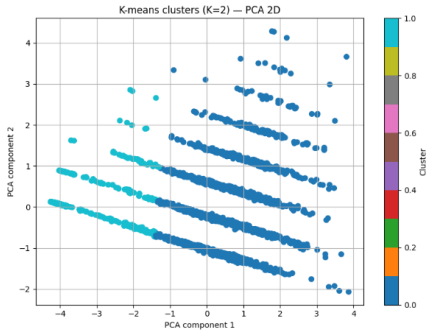
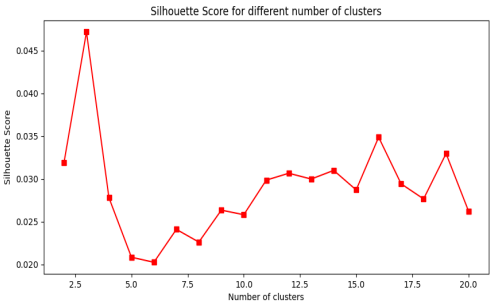
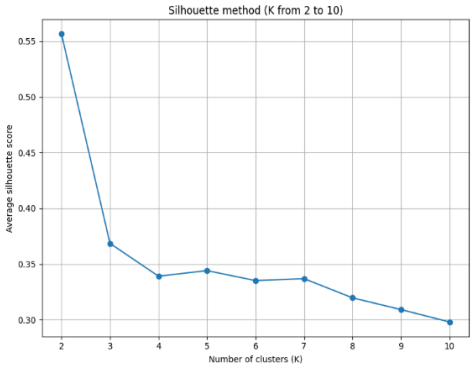
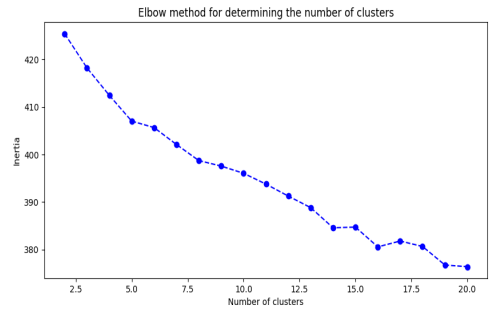
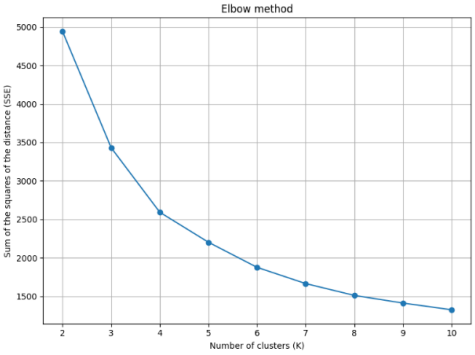
2) Analysis of clustering quality via the Silhouette Score metric,

3) Determination of the optimal number of clusters using the Elbow method.

These three levels of analysis together provide a comprehensive view of the data structure and algorithm performance in different contexts.

4.3. Visual Analysis of the Obtained Results

Table 2. *Comparative analysis of visual representations of clustering.*

TEXTUAL DATA	NUMERICAL DATA
	
	
	

In the left column of Table 2, the visual clustering results for the textual data are shown. The first image reveals that the points, representing individual legal articles, are distributed across multiple areas without clear boundaries between them. The feature space, obtained from the transformation of TF-IDF vectors, shows a high degree of overlap among terms, which is typical for legal terminology where the same words appear in different contexts. Silhouette analysis indicated an average value of only 0.035, pointing to weak internal cohesion of clusters and significant overlap between them. The graphical representation of Silhouette results confirms values in the range of 0.02–0.045, without pronounced local maxima. The Elbow method did not show a clear "elbow point" but a gradual decline

of inertia from 1200 to 850 as the number of clusters increased from K=2 to K=10. This suggests that there is no natural cluster structure in the data and that the data are semantically continuous. These results indicate that the K-Means algorithm, which assumes linear separation, is not suitable for clustering legal texts without additional semantic modeling (e.g., using Word2Vec, BERT, or topic clustering via LDA).

The right column of Table 2 shows the clustering results for the numerical dataset related to traffic accidents (attributes such as number of vehicles, road type, severity of consequences, and weather conditions). For this dataset, the PCA visualization clearly shows the formation of two compact and separated groups, which is the first indication of natural segmentation. Silhouette analysis showed a maximum value of 0.557 for K=2, which is an excellent result in clustering. Values then sharply decrease for larger K, further confirming the existence of two natural groups in the data. This is consistent with the Elbow method, which shows a characteristic "break" at K=2, where inertia sharply drops from 1040 to 650 and then stabilizes.

These results clearly demonstrate that the K-Means algorithm is highly effective in this case, as the data have clearly defined numerical relationships and low dimensionality, enabling precise separation based on Euclidean distance.

4.3. Concluding Observations of the Results

The obtained results provide the following key insights:

1) **Textual Data:** Low Silhouette metric values and the absence of a clear "elbow" confirm that K-Means fails to detect semantic similarities in legal texts. This suggests the need for more advanced techniques (e.g., BERT embeddings or topic-based clustering).

2) **Numerical Data:** High Silhouette scores and the stable drop in inertia at K=2 indicate that K-Means can be very effective in clustering well-structured data.

3) **Methodological Conclusion:** The nature of the data (textual or numerical) determines the applicability of the algorithm. Different types of data require tailored preprocessing and model selection approaches.

Overall, the presented results and analysis show that the K-Means algorithm can deliver high-quality outcomes only when input data are well-defined and linearly separable. In contrast, textual and semantically rich datasets require methodological extensions and deeper context-aware models.

5. DISCUSSION OF CLUSTERING METHODS

The analysis and clustering results clearly show that each algorithm provides a different perspective on the structure of the Criminal Code text. K-Means demonstrated its strength in quickly segmenting large datasets and clearly dividing them into thematic groups, but careful selection of the number of clusters is required – small variations in the n_clusters parameter can significantly alter the results. DBSCAN revealed natural "groups" of documents and identified several articles that do not fit into standard themes, indicat-

ing potential anomalies or articles with specific subject matter. Mean-Shift confirmed the existence of density centers in the texts, but it was slower on larger sets of TF-IDF vectors, suggesting it is better suited for detailed analysis of smaller subsets. Hierarchical clustering proved to be the most visually informative, allowing the tracking of how Criminal Code articles group together at different levels, which is useful for legal analysis and the creation of thematic maps. [10]

From the code and experiments, a practical insight can also be drawn: the key to good results lies in high-quality preprocessing and TF-IDF transformation. Without this step, even the most sophisticated algorithms could not differentiate between semantically distinct but superficially similar articles. Additionally, combining the results of multiple methods allows for the identification of both "major" themes (K-Means, hierarchical) and "unusual" or anomalous groups (DBSCAN, Mean-Shift), which has practical applications in legal analysis: it makes it easier to spot articles that deviate or may be relevant to specific cases.

This insight shows that, in practice, the choice of algorithm should not be rigid – it should be adapted to the type of data, the goal of the analysis, and the need to identify both core groups and unusual exceptions within the text.

6. RECOMMENDATIONS FOR FUTURE RESEARCH AND DIRECTIONS FOR IMPROVEMENT

The comparative analysis shows that:

– The K-Means algorithm has a clear advantage with numerical, well-structured data, providing stable and interpretable results.

– For textual data, the traditional TF-IDF + PCA approach provides only limited insight, and the results are not sufficiently stable, indicating the need to integrate semantic and deep learning models.

– Future research is recommended to experiment with different text vectorization models, as well as to combine textual and numerical data for multimodal analysis, in order to gain deeper insight into complex social or legal phenomena.

Although the study confirmed that K-Means can successfully process both types of data, the results indicate significant potential for improvement. Key directions for enhancing future research can be summarized in several areas, as outlined in Table 3:

1) Application of Advanced Techniques for Textual Data

   – Instead of the classical TF-IDF representation, it is preferable to apply deep learning models for natural language processing, such as Word2Vec, GloVe, or modern BERT models; [11]

   – These models allow the creation of vectors containing richer semantic information, which can result in clearer and more meaningful clusters, as well as better metric values, such as the Silhouette Score.

2) Combining Different Clustering Algorithms

– In addition to the K-Means algorithm, which requires a predefined number of clusters and assumes spherical group shapes, it is recommended to explore alternatives such as DBSCAN and Hierarchical Clustering;

– These algorithms often provide better results for unstructured and semantically rich data, as they can identify irregular and overlapping structures within the dataset.

3) Expansion of Datasets and Dimensionality Reduction

– For numerical data, it is recommended to include additional attributes such as weather conditions, road type, time of day, driver age, and other contextual factors. This would enable better understanding of causal relationships between variables and increase cluster accuracy.

– For textual data, expanding the corpus to include a larger number of legal articles or other normative documents can lead to more stable and statistically reliable results.

– The use of dimensionality reduction techniques such as PCA, t-SNE, or UMAP can improve visualization and the identification of patterns in high-dimensional data. [12]

Table 3. *Recommendations for improving clustering results.*

Data Type	Current Result	Recommendations for Improvement	Expected Effect
Numerical (traffic accidents)	Higher Silhouette Score (≈ 0.35–0.56), clearly formed clusters	- Expand the dataset (weather conditions, road type, time of day, driver age) - Apply DBSCAN and Hierarchical Clustering - Remove "noisy" data	More precise and richer clusters, better understanding of risk factors
Textual (legal articles)	Low Silhouette Score (≈ 0.02–0.045), overlapping clusters	- Use NLP models (Word2Vec, GloVe, BERT) - Combine multiple clustering methods - Apply dimensionality reduction (PCA, t-SNE, UMAP)	Clearer semantic relationships and more structured classification of legal concepts

7. CONCLUSION

This study conducted a comparative analysis of the application of the K-Means clustering algorithm on textual and numerical data. The results demonstrated that the nature of the data significantly affects the quality and stability of the clustering.

For textual data, clustering enabled thematic grouping of the articles of the Criminal Code, but with low Silhouette scores (0.02–0.045) and noticeable cluster overlap. This indicates the need for more precise text vector modeling and potential application of semantic models to achieve a clearer cluster structure. On the other hand, numerical data on traffic accidents exhibited stable and interpretable clustering, with an optimal number of clusters K=2 and a good Silhouette score (≈0.557). These results allow practical applications in trend monitoring, planning preventive measures, and analyzing traffic safety.

Based on the comparative analysis, it can be concluded that the K-Means algorithm provides more reliable and easily interpretable results for numerical data, whereas textual data require additional preprocessing techniques and advanced vectorization methods. Combining both approaches offers potential for integrated analyses, encompassing both the legislative framework and statistical indicators of social reality.

This work provides a foundation for further research on the application of machine learning algorithms to heterogeneous datasets and highlights the importance of method selection depending on the nature of the data and the objectives of the analysis.

FUNDING:
This research received no external funding.

INSTITUTIONAL REVIEW BOARD STATEMENT:
Not applicable.

INFORMED CONSENT STATEMENT:
Not applicable.

CONFLICTS OF INTEREST:
The author declares no conflict of interest.

REFERENCES

1] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* vol. 1, pp. 281–297, 1967.

[2] X. &. Z. H. Cheng, "Clustering Analysis for High-Dimensional Data Using K-Means and PCA," *Data Science and Engineering,* vol. 5, pp. 107–118, 2020.

[3] L. K. V. Havrlant, "A Simple Probabilistic Explanation of Term Frequency-Inverse Document Frequency (TF-IDF) Heuristic," *International Journal of General Systems,* vol. 46, pp. 27–36, 2017.

[4] X. &. X. S. Wu, "Machine Learning for Text Classification: A Survey," *International Journal of Computational Intelligence,* vol. 4, pp. 143–154, 2008.

[5] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Graphical Statistics,* vol. 1, pp. 53–65, 1987.

[6] F. L. Gewers, G. R. Ferreira, H. F. de Arruda, F. N. Silva, C. H. Comin, D. R. Amancio and L. da Costa, "Principal Component Analysis: A Natural Approach to Data Exploration," 2018. Available: https://arxiv.org/abs/1804.02502. [Accessed: 02 October 2025].

[7] G. H. Dunteman, Principal Components Analysis, SAGE Publications, 1989.

[8] I. T. Jolliffe, Principal Component Analysis, Springer, 2002.

[9] C. D. Manning, P. Raghavan and H. Schütze, Introduction to Information Retrieval, Cambridge, United Kingdom: Cambridge University Press, 2008.

[10] D. &. T. Y. Xu, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, pp. 165–193, 2015.

[11] T. C. K. C. G. &. D. J. Mikolov, "Efficient Estimation of Word Representations in Vector Space," Cornell University (arXiv), 2013.

[12] J. S. R. M. C. D. Pennington, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.