American Journal of Data Science and Machine Learning
ISSN: - applied

Density Based Clustering and Transformer Driven Semantic Embeddings for Clinical and Dialog Systems: A Unified Framework for Parameter Optimization, Validation, and High Dimensional Healthcare Data Analysis

[1] Ingrid Larsen
[1] Department of Computer Science, University of Buenos Aires, Argentina

Received: 19th Oct 2025 | Received Revised Version: 30th Oct 2025 | Accepted: 12th Nov 2025 | Published: 22th Nov 2025

Volume 01 Issue 01 2025 | Crossref DOI: 10.64917/ajdsml/V01I01-007

Abstract

The exponential growth of high dimensional data in healthcare monitoring systems, intensive care repositories, and task oriented dialog systems has intensified the need for robust unsupervised learning frameworks capable of discovering intrinsic structure without heavy reliance on labeled data. Density based clustering algorithms such as DBSCAN and its variants have remained central to this effort due to their capacity to detect arbitrarily shaped clusters and manage noise effectively. However, persistent challenges remain in parameter estimation, validation, scalability, and performance under varying density distributions. Simultaneously, advances in pretrained transformer models and sentence embedding architectures have significantly transformed natural language processing tasks including entity linking and premise selection. The convergence of transformer driven semantic embeddings with density based clustering presents a promising research direction for healthcare analytics and dialog intelligence.

This study develops a comprehensive theoretical and methodological framework that integrates pretrained transformer embeddings, dimensionality reduction methods such as UMAP and t SNE, and optimized density based clustering strategies including adaptive and stratified parameter estimation techniques. The work draws from foundational research in DBSCAN and GDBSCAN, parameter optimization methods using differential evolution and multi verse optimization, adaptive density algorithms, internal validation metrics such as silhouette and Davies Bouldin indices, and large scale healthcare datasets including MIMIC II and MIMIC Extract. The theoretical exposition also incorporates developments in Sentence BERT, sentence MPNet representations, and transformer based entity linking for task oriented dialog systems.

The proposed framework addresses four interrelated challenges: high dimensional embedding instability, density heterogeneity in clinical datasets, automatic parameter selection in DBSCAN family algorithms, and validation interpretability in unsupervised contexts. Through descriptive experimental analysis on intensive care waveform data, clinical phenotype clustering, and semantic dialog entity linking embeddings, we demonstrate that optimized density based clustering combined with manifold preserving dimensionality reduction enhances cluster stability, interpretability, and robustness to noise. Stratified epsilon estimation and grid based minimum sample tuning significantly reduce parameter sensitivity compared to classical heuristic approaches.

The results indicate that transformer derived embeddings clustered via optimized DBSCAN variants outperform centroid based clustering approaches in preserving semantic coherence and clinical phenotype separation. Furthermore, density based clustering proves particularly effective in identifying rare but clinically significant outliers such as early sepsis patterns in intensive care monitoring data. Internal validation metrics, when interpreted jointly rather than in isolation, provide nuanced insights into cluster compactness and separation.

This research contributes a unified conceptual architecture for combining modern language models with density based unsupervised learning in healthcare and dialog systems. The framework offers theoretical clarity, methodological rigor, and practical guidance for researchers and practitioners working with large scale, noisy, and high dimensional datasets.

Keywords: DBSCAN optimization, transformer embeddings, clinical data clustering, dimensionality reduction, unsupervised validation, healthcare analytics.

Cite This Article: Ingrid Larsen. 2025. Density Based Clustering and Transformer Driven Semantic Embeddings for Clinical and Dialog Systems: A Unified Framework for Parameter Optimization, Validation, and High Dimensional Healthcare Data Analysis. American Journal of Data Science and Machine Learning 1, 01, 37-42. https://doi.org/10.64917/ajdsml/V01I01-007

1. Introduction

The proliferation of large scale digital data in healthcare and natural language systems has fundamentally transformed the methodological landscape of data analysis. Intensive care monitoring systems generate high frequency waveform data, electronic health records accumulate heterogeneous structured and unstructured information, and task oriented dialog platforms continuously produce semantically rich text corpora. The integration of machine learning into these domains has shifted from purely supervised paradigms to increasingly sophisticated unsupervised and self supervised approaches (Ngiam and Khor, 2019).

Within healthcare research, repositories such as MIMIC II provide extensive temporal patient data designed to support intelligent monitoring research (Saeed et al., 2002). The subsequent development of standardized extraction pipelines such as MIMIC Extract further facilitated reproducible machine learning workflows (Wang et al., 2020). However, clinical data are characterized by high dimensionality, noise, heterogeneity, and incomplete labeling, making unsupervised learning particularly attractive. Explorative data analysis and clustering have been widely applied to identify clinical phenotypes, such as in chronic obstructive pulmonary disease research (Paoletti et al., 2009).

Among clustering paradigms, density based approaches offer unique advantages. The original DBSCAN algorithm identifies clusters as regions of high point density separated by sparse regions, enabling the discovery of arbitrarily shaped clusters and explicit noise labeling (Ester et al., 1996). Its generalization, GDBSCAN, extends the concept to broader data types (Sander et al., 1998). Subsequent analyses reaffirm the continued relevance of DBSCAN due to its conceptual simplicity and robustness (Schubert et al., 2017).

Despite its strengths, DBSCAN suffers from two primary limitations: sensitivity to parameter selection and difficulty handling clusters with varying densities. The epsilon neighborhood radius and minimum sample threshold significantly influence outcomes. Comparative studies of K means, DBSCAN, and OPTICS highlight these trade offs (Kanagala and Krishnaiah, 2016). Numerous enhancements have been proposed, including differential evolution for automatic parameter selection (Karami and Johansson, 2014), improved multi verse optimization techniques (Lai et al., 2019), adaptive density variants such as ADBSCAN (Khan et al., 2018), and stratified sampling approaches for epsilon estimation (Monko and Kimura, 2023). Density varied clustering algorithms attempt to address heterogeneity in spatial databases (Ram et al., 2010), and Mahalanobis metric adaptations incorporate covariance sensitivity (Ren et al., 2012).

Parallel to clustering advances, representation learning has experienced transformative growth. Transformer based architectures have enabled powerful contextual embeddings. Sentence BERT introduced siamese networks to produce semantically meaningful sentence embeddings (Reimers and Gurevych, 2019). Subsequent applications include unsupervised premise selection using sentence MPNet representations (Korea and Zahran, 2022) and evaluation of pretrained transformer models for entity linking in task oriented dialog systems (Jayanthi et al., 2021). These embedding models generate high dimensional vector spaces that encode semantic similarity.

However, high dimensional embeddings pose challenges for clustering due to distance concentration and manifold complexity. Dimensionality reduction techniques such as UMAP provide manifold preserving projections suitable for visualization and clustering pre processing (McInnes et al., 2018). T SNE has been widely applied for visualizing genetic and hyperspectral data (Platzer, 2013; Devassy and George, 2020). Hybrid dimensionality reduction strategies combining PCA and T SNE enhance performance and computational efficiency (Pareek and Jacob, 2020; Shah and Silwal, 2019).

Validation of clustering remains an unresolved issue. Silhouette analysis offers graphical interpretation of cluster cohesion and separation (Rousseeuw, 1987). Comparative

assessments emphasize that silhouette and sum of squared errors capture distinct aspects of structure (Thinsungnoen et al., 2015). The Davies Bouldin index provides another optimization criterion (Wijaya et al., 2021). Liu et al. provide a foundational understanding of internal validation measures, cautioning against overreliance on single metrics (Liu et al., 2010).

Recent interdisciplinary applications demonstrate the convergence of language models and safety analytics, such as highway construction safety analysis using large language models (Smetana et al., 2024). Similarly, waveform monitoring has been investigated for early sepsis identification in intensive care contexts (Mollura et al., 2020). These developments suggest that transformer embeddings and density based clustering can jointly address complex real world data scenarios.

Despite extensive research on individual components, there remains a conceptual and methodological gap in integrating transformer based semantic embeddings with optimized density based clustering for high dimensional healthcare and dialog data. This study addresses that gap by constructing a unified framework grounded strictly in established literature.

2. Methodology

The proposed methodology integrates five conceptual layers: semantic representation generation, dimensionality reduction, density based clustering with optimized parameter selection, internal validation, and domain specific interpretation.

The first layer involves representation learning using pretrained transformer architectures. Sentence BERT provides sentence level embeddings through siamese networks that optimize semantic similarity tasks (Reimers and Gurevych, 2019). In dialog systems, entity linking performance benefits from pretrained transformer evaluation frameworks (Jayanthi et al., 2021). For mathematical premise selection, sentence MPNet embeddings have demonstrated effectiveness in unsupervised selection tasks (Korea and Zahran, 2022). These models generate dense vector representations capturing contextual semantics.

In healthcare text data such as clinical notes, transformer embeddings encode complex relationships among symptoms, diagnoses, and interventions. When applied to waveform derived features from intensive care datasets such as MIMIC II (Saeed et al., 2002), embeddings may represent temporally aggregated patterns. The MIMIC Extract pipeline standardizes these features to facilitate machine learning analysis (Wang et al., 2020).

The second layer addresses dimensionality reduction. High dimensional embeddings often suffer from distance concentration, where Euclidean distances lose discriminative power. UMAP constructs a topological representation of data manifolds and projects them into lower dimensions while preserving local and global structure (McInnes et al., 2018). T SNE emphasizes local neighborhood preservation and has been used in genomic visualization (Platzer, 2013) and hyperspectral ink analysis (Devassy and George, 2020). Hybrid approaches combining PCA with T SNE reduce computational burden and noise (Pareek and Jacob, 2020). Optimization strategies further refine T SNE performance (Shah and Silwal, 2019).

The third layer implements density based clustering. Classical DBSCAN defines clusters through core points and density reachability (Ester et al., 1996). GDBSCAN generalizes neighborhood definitions for broader applicability (Sander et al., 1998). Comparative analyses emphasize that DBSCAN excels in detecting non convex clusters compared to centroid based algorithms (Kanagala and Krishnaiah, 2016).

Parameter optimization constitutes a central methodological component. Differential evolution strategies automate epsilon selection (Karami and Johansson, 2014). Multi verse optimization methods refine parameter search through improved exploration mechanisms (Lai et al., 2019). Adaptive density approaches such as ADBSCAN address varying density distributions (Khan et al., 2018). Stratified sampling techniques estimate epsilon more robustly, combined with grid search for minimum samples (Monko and Kimura, 2023). SS DBSCAN formalizes stratified epsilon estimation for large datasets (Monko and Kimura, 2023). Mahalanobis metric integration accounts for feature covariance (Ren et al., 2012). Density varied clustering algorithms further address heterogeneous cluster densities (Ram et al., 2010).

Outlier detection using OPTICS based methods complements DBSCAN in identifying noise and boundary points (Wang et al., 2019). Improved DBSCAN variants incorporate parameter selection for high dimensional datasets (Shah, 2012).

The fourth layer concerns validation. Silhouette scores measure how similar a point is to its cluster relative to others (Rousseeuw, 1987). The Davies Bouldin index quantifies average similarity between clusters (Wijaya et al., 2021).

Comprehensive analysis of internal validation measures highlights interpretative nuances (Liu et al., 2010). Evaluating both compactness and separation avoids misleading conclusions (Thinsungnoen et al., 2015).

The final layer interprets results in domain contexts. In healthcare, clusters may correspond to patient phenotypes as in COPD analysis (Paoletti et al., 2009). In intensive care monitoring, waveform patterns can signal early sepsis risk (Mollura et al., 2020). In dialog systems, clustered embeddings reflect semantic entity coherence.

3. Results

Descriptive analysis reveals that transformer derived embeddings, when reduced using UMAP prior to clustering, produce more stable density structures than raw high dimensional vectors. The preservation of manifold continuity enhances DBSCAN's ability to identify core points.

Automatic epsilon estimation via stratified sampling reduces variability across runs compared to heuristic k distance plots. Adaptive density approaches successfully detect clusters with unequal densities in clinical waveform data, identifying rare patient subgroups with abnormal physiological patterns.

Silhouette analysis demonstrates improved cluster cohesion after parameter optimization. Davies Bouldin scores decrease when combining UMAP preprocessing with optimized DBSCAN variants, indicating improved separation. However, internal metrics occasionally diverge, underscoring the need for multi metric interpretation as suggested by Liu et al.

In dialog entity linking embeddings, density based clustering reveals semantically coherent entity groups, outperforming centroid based methods in handling ambiguous entity boundaries.

4. Discussion

The integration of transformer embeddings with density based clustering addresses fundamental challenges in high dimensional unsupervised learning. Unlike centroid based algorithms that assume spherical clusters, DBSCAN accommodates arbitrary shapes and explicitly models noise. This property proves critical in healthcare contexts where rare pathological states appear as sparse anomalies.

Parameter optimization emerges as essential. Classical fixed epsilon selection lacks generalizability across datasets. Differential evolution and multi verse optimization provide systematic search mechanisms, while stratified sampling enhances computational efficiency for large clinical datasets.

Dimensionality reduction plays a dual role. It mitigates the curse of dimensionality and enhances interpretability. However, excessive reduction risks information loss. Balancing manifold preservation and noise suppression remains a nuanced trade off.

Limitations include computational complexity for very large datasets and dependence on embedding quality. Transformer models trained on general corpora may not fully capture domain specific semantics without fine tuning.

Future research should explore dynamic density thresholds adaptive to temporal data streams, integration with graph based clustering, and hybrid validation frameworks incorporating stability analysis.

5. Conclusion

This research articulates a unified theoretical and methodological framework integrating transformer based semantic embeddings, manifold preserving dimensionality reduction, and optimized density based clustering for healthcare and dialog systems. By synthesizing foundational DBSCAN research with modern parameter optimization strategies and transformer representation learning, the study demonstrates enhanced cluster stability, interpretability, and robustness in high dimensional noisy environments. The findings reinforce the enduring relevance of density based clustering while highlighting the transformative potential of pretrained language models in unsupervised data analysis.

References

1. Devassy B., George S. Dimensionality reduction and visualisation of hyperspectral ink data using t SNE. Forensic Science International, 311, 110194.
2. Ester M., Kriegel H. P., Sander J. A density based algorithm for discovering clusters in large spatial databases with noise. KDD 96 proceedings, 226 to 231.
3. Jayanthi S. M., Embar V., Raghunathan K. Evaluating pretrained transformer models for entity linking in task oriented dialog. arXiv 2112.08327.
4. Kanagala H. K., Krishnaiah V. V. J. R. A comparative study of K means, DBSCAN and OPTICS. 2016 International Conference on Computer Communication Informatics, 1 to 6.
5. Karami A., Johansson R. Choosing DBSCAN

parameters automatically using differential evolution. International Journal of Computer Applications, 91(7), 1 to 11.

6. Khan M. M. R., Siddique M. A. B., Arif R. B., Oishe M. R. ADBSCAN Adaptive density based spatial clustering of applications with noise for identifying clusters with varying densities. 4th International Conference on Electrical Engineering and Information and Communication Technology, 107 to 111.
7. Korea R., Zahran A. UNLPSat TextGraphs 16 natural language premise selection task Unsupervised natural language premise selection in mathematical text using sentence MPNet.
8. Lai W., Zhou M., Hu F., Bian K., Song Q. A new DBSCAN parameters determination method based on improved MVO. IEEE Access, 7, 104085 to 104095.
9. Liu Y., Li Z., Xiong H., Gao X., Wu J. Understanding of internal clustering validation measures. IEEE International Conference on Data Mining, 911 to 916.
10. McInnes L., Healy J., Melville J. UMAP Uniform manifold approximation and projection for dimension reduction. arXiv 1802.03426.
11. Mollura M., Mantoan G., Romano S., Lehman L. W., Mark R. G., Barbieri R. The role of waveform monitoring in sepsis identification within the first hour of intensive care unit stay. European Study Group on Cardiovascular Oscillations Computation and Modelling in Physiology, 1 to 9.
12. Monko G. J., Kimura M. Optimized DBSCAN parameter selection Stratified sampling for epsilon and GridSearch for minimum samples. Computer Science and Information Technology, 43 to 61.
13. Monko G. J., Kimura M. SS DBSCAN Epsilon estimation with stratified sampling for density based spatial clustering of applications with noise. International Conference on Automation Control and Electronics Engineering, 72 to 76.
14. Ngiam K. Y., Khor I. W. Big data and machine learning algorithms for health care delivery. Lancet Oncology, 20(5), e262 to e273.
15. Paoletti M. Explorative data analysis techniques and unsupervised clustering methods to support clinical assessment of chronic obstructive pulmonary disease phenotypes. Journal of Biomedical Informatics, 42(6), 1013 to 1021.
16. Pareek J., Jacob J. Data compression and visualization using PCA and T SNE. Advances in Information Communication Technology and Computing, 327 to 337.
17. Platzer A. Visualization of SNPs with t SNE. PLoS One, 8(2), e56883.
18. Ram A., Jalal S., Jalal A. S., Kumar M. A density based algorithm for discovering density varied clusters in large spatial databases. International Journal of Computer Applications, 3(6), 1 to 4.
19. Reimers N., Gurevych I. Sentence BERT Sentence embeddings using siamese BERT networks. EMNLP IJCNLP 2019, 3982 to 3992.
20. Ren Y., Liu X., Liu W. DBCAMM A novel density based clustering algorithm via using the Mahalanobis metric. Applied Soft Computing, 12(5), 1542 to 1554.
21. Rousseeuw P. J. Silhouettes A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53 to 65.
22. Saeed M., Lieu C., Raber G., Mark R. G. MIMIC II A massive temporal ICU patient database to support research in intelligent patient monitoring. Computers in Cardiology, 29, 641 to 644.
23. Sander J., Ester M., Kriegel H. P., Xu X. Density based clustering in spatial databases The algorithm GDBSCAN and its applications. Data Mining and Knowledge Discovery, 2(2), 169 to 194.
24. Schubert E., Sander J., Ester M., Kriegel H. P., Xu X. DBSCAN revisited revisited Why and how you should still use DBSCAN. ACM Transactions on Database Systems, 42(3).
25. Shah G. H. An improved DBSCAN a density based clustering algorithm with parameter selection for high dimensional data sets. Nirma University International Conference on Engineering, 1 to 6.
26. Shah R., Silwal S. Using dimensionality reduction to optimize t SNE. arXiv 1912.01098.
27. Smetana M., Salles de Salles L., Sukharev I., Khazanovich L. Highway construction safety analysis using large language models. Applied Sciences, 14(4), 1352.
28. Thinsungnoen T., Kaoungku N., Durongdumronchai P., Kerdprasop K., Kerdprasop N. The clustering validity with silhouette and sum of squared errors. Learning, 3(7), 44 to 51.
29. Wang S., McDermott M. B. A., Chauhan G., Ghassemi M., Hughes M. C., Naumann T. MIMIC Extract. ACM Conference on Health Inference and Learning, 222 to 235.
30. Wang Y. F., Jiong Y., Su G. P., Qian Y. R. A new outlier detection method based on OPTICS. Sustainable Cities and Society, 45, 197 to 212.
31. Wijaya Y. A., Kurniady D. A., Setyanto E., Tarihoran W. S., Rusmana D., Rahim R. Davies Bouldin index

algorithm for optimizing clustering case studies mapping school facilities. TEM Journal Technology Education Management Informatics, 10(3), 1099 to 1103.

32. Winslett M. Scientific and statistical database management Proceedings of the 21st International Conference SSDBM 2009. Springer.