# Clustering Medical Transcriptions Using K-Means

Said Salloum
School of Science, Engineering, and Environment
University of Salford
Manchester, UK
salloum78@live.com

Dina Tahat
Al Ain University, Applied sociology department, UAE
dina.tahat@aau.ac.ae

Raghad Alfaisal
Faculty of Computing and Meta-Technology, Universiti Pendidikan Sultan Idris, Tanjung Malim, Malaysia
raghad.alfaisal81@gmail.com

Khalaf Tahat
Media & Creative Industries, United Arab Emirates University and Yarmouk University, Jordan
k.tahat@uaeu.ac.ae

Ayham Salloum
College of Medicine, University of Sharjah, Sharjah, UAE
ssalloum1978@gmail.com

***Abstract*—The clustering of medical transcriptions is an essential task for the categorization and summarization of large volumes of medical records. This paper explores the efficacy of k-means clustering, a well-known unsupervised machine learning algorithm, to discern patterns and segregate medical transcriptions into distinct clusters. We processed a dataset comprising various medical reports, systematically cleaning and preparing the text for analysis. By employing a Term Frequency-Inverse Document Frequency (TF-IDF) approach, we converted the textual data into a vectorized format amenable to machine learning methods. Subsequent dimensionality reduction through Principal Component Analysis (PCA) facilitated the visualization and interpretation of the high-dimensional data in two-dimensional space. The k-means algorithm was then applied, revealing five distinct clusters. Each cluster was characterized by examining the prevalence of key terms, uncovering thematic consistencies that may correspond to particular medical procedures or specialties. The resulting clusters demonstrate the algorithm's potential to automatically categorize medical documentation in a way that mirrors clinical relevance, thereby providing a foundation for improved information management systems in healthcare settings.**

**Keywords—Dimensionality Reduction, K-Means Clustering, Medical Transcriptions, Unsupervised Learning.**

## I. Introduction

The proliferation of digital records in healthcare has introduced both challenges and opportunities in medical data analysis. Among these records, medical transcriptions play a crucial role as they capture detailed narratives of patient encounters, treatments, and outcomes. Efficient categorization and retrieval of these transcriptions are paramount for enhancing patient care and supporting healthcare professionals in their decision-making processes.

Despite their importance, medical transcriptions are often underutilized due to their unstructured format, which poses significant hurdles for text analysis and data retrieval [1]. Machine learning offers sophisticated methods for analyzing and deriving meaningful patterns from such data. Among these methods, unsupervised machine learning algorithms, especially clustering techniques, have shown promise in discerning inherent groupings in text data without predefined labels [2].

K-means clustering is one of the most widely employed unsupervised learning algorithms due to its simplicity and effectiveness in various domains [3]. It operates by partitioning n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. In the domain of text mining, k-means has been successfully applied to group documents by themes or subjects, facilitating information retrieval and organization [4]–[6].

In this paper, we aim to leverage k-means clustering to categorize medical transcriptions. By converting the textual content into a numerical format through Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, we prepare the data for the application of k-means. This process allows us to identify natural clusters in the dataset, which we hypothesize to correspond with various medical specialties or types of patient encounters, mirroring the taxonomy used by healthcare professionals.

This research contributes to the field of medical data analysis by:

- Demonstrating the application of k-means clustering to medical transcriptions, an area that poses significant challenges due to the specialized and diverse nature of the content.
- Introducing a comprehensive methodology for preprocessing and vectorizing medical text data to identify distinct themes and patterns.
- Presenting a novel perspective on the categorization of medical documents that aligns with clinical workflows and specialty areas, which may facilitate enhanced retrieval and analysis systems.

The remainder of this paper is structured as follows: Section 2 outlines the related work and theoretical background supporting the use of k-means in text clustering. Section 3 describes the dataset and details the methods employed for data preprocessing, vectorization, and clustering. Section 4 presents the results of the clustering process, including an analysis of the terms most characteristic of each cluster. Section 5 discusses the

979-8-3503-5469-0/24/$31.00 ©2024 IEEE

implications of our findings, the potential applications in healthcare settings, and the limitations of our study. Finally, Section 6 concludes the paper with a summary of the research and suggestions for future work in the area.

## II. Methodology

### A. Dataset Description

The dataset employed in this study comprises medical transcriptions publicly available from the Kaggle dataset repository [7]. These transcriptions encapsulate a wide range of medical reports documented by healthcare professionals, encompassing various medical terminologies and specialties. The dataset is representative of the typical documentation found in electronic health record systems, which include patient histories, diagnostic findings, and treatment courses.

### B. Data Preprocessing

Transcription data inherently contains noise and irrelevant information that can skew analysis results. Therefore, a rigorous preprocessing routine was implemented to sanitize the data [8], [9]. This process involved converting all text to lowercase to ensure uniformity and removing non-alphanumeric characters that are typically irrelevant for text analysis [10]. Additionally, stopwords—words in the English language that carry minimal individual meaning, such as "the," "is," and "and"—were excluded from the dataset to better concentrate on medically pertinent terms [11], [12].

### C. Vectorization

To analyze the textual data quantitatively, we converted the preprocessed transcriptions into a TF-IDF matrix. This method weighs the terms within the document against their frequency across the entire corpus, diminishing the impact of commonly occurring words while emphasizing unique terms [13]. The TF-IDF vectorization is a widely recognized approach in natural language processing for preparing text for machine learning applications [14], [15].

### D. Dimensionality Reduction

Given the high-dimensional nature of TF-IDF vectors, Principal Component Analysis (PCA) was utilized to reduce the dimensionality of the dataset [16]. PCA was performed to distill the vectors into two principal components, which capture the most significant variance within the data while enabling visualization in a two-dimensional space. This technique simplifies the complexity of the dataset while retaining its structural integrity [17], [18].

### E. Clustering

K-means clustering was selected for its efficacy in partitioning data into k distinct clusters by minimizing the within-cluster variance [19], [20]. The optimal number of clusters, k, was determined to be 5, a decision based on the dataset's inherent characteristics and preliminary analysis [21], [22]. Each transcription was then assigned to the nearest cluster centroid, based on the Euclidean distance in the reduced feature space.

## III. Findings and Discussion

The application of k-means clustering to the PCA-reduced TF-IDF matrix of medical transcriptions successfully partitioned the dataset into five distinct clusters. These clusters were meticulously analyzed to identify the most frequent terms, which were then used to ascertain the predominant themes and possible medical specializations they may represent.

### A. Cluster Analysis and Keywords

- Cluster 0 was characterized by terms indicative of surgical or invasive procedures. Frequent references to 'patient placement', 'incisions', and 'anesthesia use' suggest a focus on operative reports and perioperative care. The specificity of terms like 'incision' points to a cluster that is likely associated with surgical transcripts, a critical area in medical documentation.
- Cluster 1 appeared to encapsulate general clinical encounters. Terms such as 'patient procedures', 'pain', and 'discharge' are common in a variety of clinical settings, ranging from emergency medicine to inpatient care. The presence of terms related to 'pain' and 'discharge' might denote documentation concerning patient assessments and treatment plans.
- Cluster 2 predominantly contains terminology associated with radiology or diagnostic imaging. The prominence of 'MRI' and 'CT scans' indicates that the transcripts in this cluster pertain to diagnostic imaging reports, a vital component in contemporary diagnostics.
- Cluster 3 is distinct in its focus on patient history and ongoing care, as reflected by the prevalence of terms such as 'history', 'pain management', and 'medication dosages'. This cluster may be indicative of internal medicine or family practice specialties, where comprehensive patient histories and long-term treatment strategies are emphasized.
- Cluster 4 is enriched with terms such as 'artery', 'coronary', 'aortic valve', and 'stenosis', which are strongly related to cardiovascular procedures. The technical nature of the terms, including 'catheter' and 'french', suggests that these transcriptions are from cardiology departments and related interventional procedures.

### B. *Visualization Interpretation*

Figure 1 presents the two-dimensional PCA visualization of the medical transcription data, post-clustering. Each cluster is color-coded, allowing for the visual inspection of the grouping efficacy. The dispersion of data points indicates the natural tendencies of transcriptions to aggregate based on their textual content, while also highlighting the potential overlaps between different medical domains.

Figure 2 delineates the distribution of transcriptions across the five clusters. The size disparity among clusters provides insights into the dataset's composition and the relative frequency of various types of medical reports. Larger clusters may indicate more common types of transcriptions or broader categories, whereas smaller clusters might correspond to more specialized or less frequent types of medical documentation.

The delineated clusters reveal a notable division of medical transcription data which corresponds well with known medical specializations and report types. These findings underscore the capacity of unsupervised learning techniques like k-means clustering to structure and categorize unlabelled textual data in a meaningful manner.

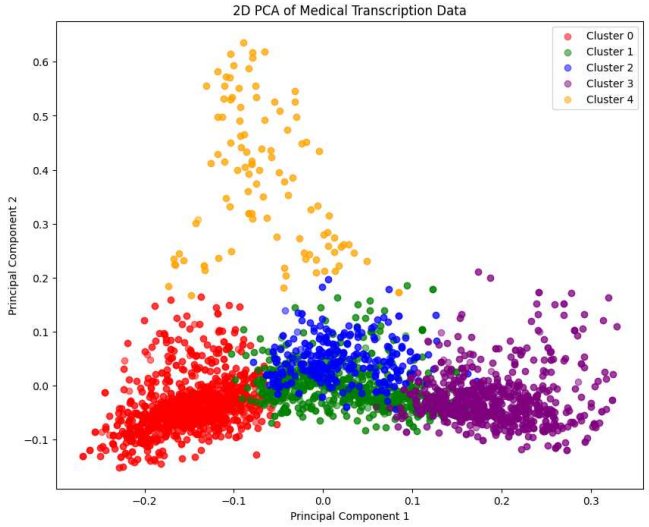


**Fig. 1.** 2D PCA visualization of the clustered medical transcriptions

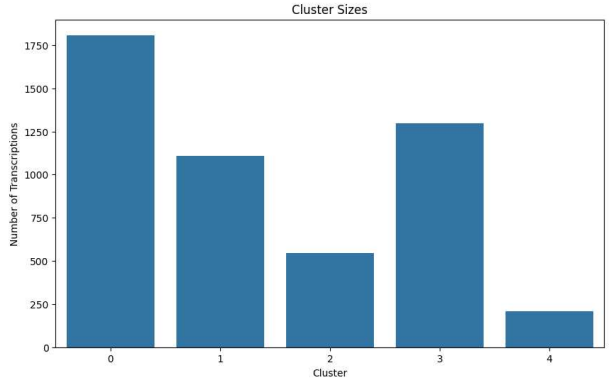


**Fig. 2.** Distribution of the number of transcriptions across the five clusters.

## I. Discussion

The results of this study affirm the potential of k-means clustering to methodically organize medical transcriptions into significant categories. The clusters formed through the analysis reflect a strong alignment with distinct medical domains, such as cardiology, radiology, and general patient care, demonstrating the algorithm's capability to recognize and differentiate between the subtle variances in medical language utilized in diverse specialties [23].

Particularly noteworthy is the distinct lexical grouping within each cluster, which suggests that k-means clustering is proficient in identifying thematic consistencies in medical documents. For instance, Cluster 4's focus on terms related to cardiovascular procedures strongly corresponds with cardiological specializations, thereby implying that the algorithm may have practical applications in assisting healthcare institutions in sorting and routing medical documentation [24].

However, the study is not without limitations. The k-means algorithm assumes spherical clusters and does not perform well with clusters of different shapes and densities [25]. Furthermore, the choice of k, while informed by preliminary analysis, remains somewhat arbitrary and could be optimized through more sophisticated methods like silhouette analysis [26].

Future research should explore the integration of k-means clustering with other natural language processing techniques, such as named entity recognition or topic modeling, to refine the categorization process [27]. Additionally, validating the clusters against external benchmarks or medical classification systems would help establish the clinical relevancy of the observed groupings [28].

## II. Coclusion

This investigation into the use of k-means clustering on medical transcription data has demonstrated the technique's robustness in generating meaningful categorizations of textual data. The discerned clusters mirror known medical specializations, underscoring the utility of this unsupervised learning approach in enhancing the organization of medical records. The implications of this research extend beyond mere data categorization. By facilitating more efficient record management, k-means clustering could conceivably contribute to improved clinical decision-making and patient care, as healthcare providers could access categorized patient information with greater ease and accuracy [29]. Additionally, the clustering of medical documents can aid in the development of specialized automated tools for medical record summarization and retrieval, thus streamlining clinical workflows [30]. In summary, the application of k-means clustering presents a promising avenue for the future of medical data analysis. It is anticipated that with further validation and integration with other computational techniques, this method could significantly enhance the efficiency and effectiveness of healthcare services.

## References


[1] H. Shatkay and R. Feldman, "Mining the biomedical literature in the genomic era: an overview," *J. Comput. Biol.*, vol. 10, no. 6, pp. 821–855, 2003.

[2] M. de F. O. Baffa, N. S. Schaadt, F. Feuerhake, and T. M. Deserno, "Unsupervised deep learning for clustering tumor subcompartments in histopathological images of non-small cell lung cancer," in *Medical Imaging 2024: Imaging Informatics for Healthcare, Research, and Applications*, 2024, vol. 12931, pp. 221–228.

[3] K. P. Sinaga and M.-S. Yang, "Unsupervised K-means clustering algorithm," *IEEE access*, vol. 8, pp. 80716–80727, 2020.

[4] S. A. Salloum, A. Q. AlHamad, M. Al-Emran, and K. Shaalan, *A survey of Arabic text mining*, vol. 740. 2018.

[5] F. Almatrooshi, S. Alhammadi, S. A. Salloum, and K. Shaalan, "Text and web content mining: a systematic review," in *International Conference on Emerging Technologies and Intelligent Systems*, 2021, pp. 79–87.

[6] M. C. Habes M., Alghizzawi M., Salloum S.A., M. Habes, M. Alghizzawi, S. A. Salloum, and C. Mhamdi, "Effects of Facebook Personal News Sharing on Building Social Capital in Jordanian Universities.," *Al-Emran M., Shaalan K., Hassanien A. Recent Adv. Intell. Syst. Smart Appl. Stud. Syst. Decis. Control. vol 295. Springer, Cham*, vol. 295, no. June 2020, pp. 653–670, 2021, doi: 10.1007/978-3-030-47411-9_35.

[7] "Medical Transcriptions," *Kaggle*, 2019. https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions.

[8] A. Kathuria, A. Gupta, and R. K. Singla, "A review of tools and techniques for preprocessing of textual data," *Comput. Methods Data Eng. Proc. ICMDE 2020, Vol. 1*, pp. 407–422, 2021.

[9] S. A. Salloum, M. Al-Emran, S. Abdallah, and K. Shaalan, "Analyzing the Arab Gulf Newspapers Using Text Mining Techniques," in *International Conference on Advanced Intelligent Systems and Informatics*, 2017, pp. 396–405, doi: 10.1007/978-3-319-64861-3_37.

[10] S. Moon, S. Pakhomov, J. Ryan, and G. B. Melton, "Automated non-alphanumeric symbol resolution in clinical texts," in *AMIA Annual Symposium Proceedings*, 2011, vol. 2011, p. 979.

[11] S. Sarica and J. Luo, "Stopwords in technical language processing," *PLoS One*, vol. 16, no. 8, p. e0254937, 2021.

[12] S. A. Salloum, M. Al-Emran, and K. Shaalan, "A Survey of Lexical Functional Grammar in the Arabic Context," *Int. J. Com. Net. Tech*, vol. 4, no. 3, 2016.

[13] X. Yang, K. Yang, T. Cui, M. Chen, and L. He, "A study of text vectorization method combining topic model and transfer learning," *Processes*, vol. 10, no. 2, p. 350, 2022.

[14] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, "A survey on text classification algorithms: From text to predictions," *Information*, vol. 13, no. 2, p. 83, 2022.

[15] M. Alshurideh, B. H. Al Kurdi, H. M. Alzoubi, and S. Salloum, *The Effect of Information Technology on Business and Marketing Intelligence Systems*. Springer, 2023.

[16] I. T. Jolliffe, "Principal Component Analysis," *SpringerSeries Stat. google Sch.*, vol. 2, pp. 903–995, 2002.

[17] G. T. Reddy *et al.*, "Analysis of dimensionality reduction techniques on big data," *Ieee Access*, vol. 8, pp. 54776–54788, 2020.

[18] S. Salloum, K. Tahat, D. Tahat, A. Mansoori, and R. Alfaisal, "Delving Into the Security & Privacy of the Metaverse Matrix," in *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2023, pp. 1–5.

[19] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, vol. 1, no. 14, pp. 281–297.

[20] S. A. Salloum, M. Al-Emran, and K. Shaalan, "Mining Text in News Channels: A Case Study from Facebook," *Int. J. Inf. Technol. Lang. Stud.*, vol. 1, no. 1, pp. 1–9, 2017.

[21] D. Marutho, S. H. Handaka, and E. Wijaya, "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news," in *2018 international seminar on application for technology of information and communication*, 2018, pp. 533–538.

[22] K. Tahat, A. Mansoori, D. N. Tahat, M. Habes, and S. Salloum, "Leveraging Soft Power: A Study of Emirati Online Journalism Through Arabic Topic Modeling," in *International Conference on Business and Technology*, 2023, pp. 13–20.

[23] M. I. Pramanik, R. Y. K. Lau, M. A. K. Azad, M. S. Hossain, M. K. H. Chowdhury, and B. K. Karmaker, "Healthcare informatics and analytics in big data," *Expert Syst. Appl.*, vol. 152, p. 113388, 2020.

[24] Ł. Ledziński and G. Grześk, "Artificial intelligence technologies in cardiology," *J. Cardiovasc. Dev. Dis.*, vol. 10, no. 5, p. 202, 2023.

[25] Q. An, S. Rahman, J. Zhou, and J. J. Kang, "A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities and challenges," *Sensors*, vol. 23, no. 9, p. 4178, 2023.

[26] G. Ogbuabor and F. N. Ugwoke, "Clustering algorithm for a healthcare dataset using silhouette score value," *AIRCC's Int. J. Comput. Sci. Inf. Technol.*, pp. 27–37, 2018.

[27] K. Raja and S. Jonnalagadda, "Natural Language Processing and Data Mining for Clinical Text.," *Healthc. Data Anal.*, vol. 36, p. 219, 2015.

[28] M. C. Massi, F. Ieva, and E. Lettieri, "Data mining application to healthcare fraud detection: a two-step unsupervised clustering method for outlier detection with administrative databases," *BMC Med. Inform. Decis. Mak.*, vol. 20, pp. 1–11, 2020.

[29] C. Castaneda *et al.*, "Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine," *J. Clin. Bioinforma.*, vol. 5, pp. 1–16, 2015.

[30] R. Pivovarov and N. Elhadad, "Automated methods for the summarization of electronic health records," *J. Am. Med. Informatics Assoc.*, vol. 22, no. 5, pp. 938–947, 2015.