Segmenting and Classifying Clinical Notes

H.D. Fernando
University of Tennessee, Knoxville, TN 37996, USA

Corresponding Author Email: hfernand@vols.utk.edu

Electronic Health Records (EHRs) contain large amounts of clinical notes, often written in different forms. These written records contain critical information that is important for patient care. However, their unstructured nature in the documents create significant difficulties to extract meaningful information. Clinicians constantly fail to find and interpret important information, specifically when these note documents are filled by multiple people from different departments. This lack of structure and semantic integration can cause significant problems in clinical care and their outcomes. Some methods try to classify note segments by using supervised classification methods, but they rely on manual annotation and doing that requires a large amount of time.

In this project, my work addresses this problem by using supervised/unsupervised methods for classifying and clustering clinical note segments using the publicly available MIMIC-III dataset. This dataset contains multiple note categories including note categories like “Radiology”, “Nursing/other”, “Discharge summary”, “ECG” and many more. Some of these categories have good segment titles and some doesn’t. For this project we have decide to only use “Nursing/other” category to overcome time and resource limitations. Typically, a single note contains multiple segments mainly patient information, medicine, assessment, pain info, plan and many more. These types of segments used in a note depends on the note category.

“Nursing/other” category data contains 822,497 unique notes and multiple records for a single patient also contains in these note data. We were able to segment “Nursing/other” note data into 2.3 millions of segmented text over 2121 segment types. The raw “Nursing/other” contains multiple segment types and number of segment type per note varies. We have used expressions-based NLP techniques to segment the text notes. First, we have tried to find subsections by following special characters. Then we have cleaned these data by removing unwanted white spaces, dates and removing unwanted headings. Also, we have made all the numbers into static number to avoid numeric variation in the data, this helps when clustering similar sections. Finally, we applied same cleaning to the titles and added all the data into a single data frame. We have removed empty titles and selected the titles that have more than 20 segment data. To see the difference when we include the title in the segment body data, we have created a separate column which contains segment body and title together.

After preprocessing and cleaning, we built a structured labeled note segment dataset containing 2.3 million segments across 2,121 unique titles. This dataset forms the foundation for downstream modeling. To analyze these segments, we extracted features using two complementary text embedding approaches (Figure 1).

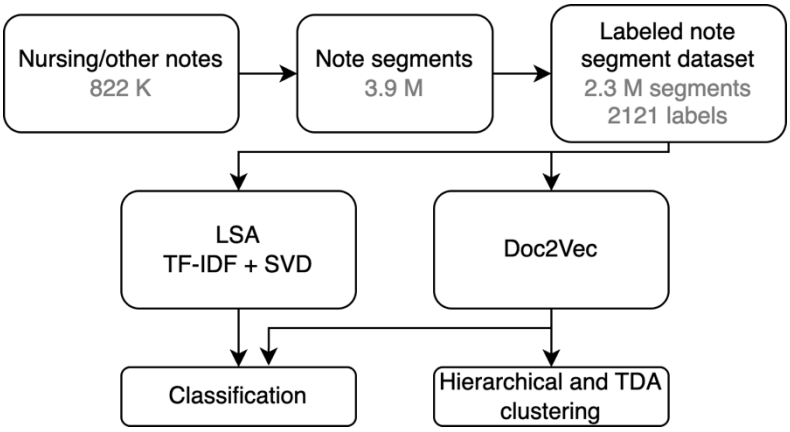


Figure 1: Project Pipeline

TF-IDF + SVD was used to represent note segments as sparse vectors and reduce dimensionality. These features were used for training a neural classification model to predict segment titles based on their body text. In parallel, we used Doc2Vec, a dense embedding technique that learns distributed representations of text segments. These embeddings were used for unsupervised clustering of note segment labels.

For clustering of segment label types, we applied Hierarchical Agglomerative Clustering and Mapper Algorithm to group segment labels based on average embedding similarity. Topological Data Analysis (TDA) using the Mapper algorithm, which constructs a network from the embedding space, helping visualize and identify semantically related labels. This dual strategy of classification and clustering enables both segment-level prediction and label-space organization, offering a scalable approach to improve clinical note navigation and interpretation.

We evaluated the performance of our classification model using F1 scores across different segment types. Segment types were sorted by frequency, with the most common labels on the left. Each box in the plot represents the distribution of F1 scores for 100 segment types grouped by frequency. The results show that high-frequency segment types tend to achieve higher and more stable F1 scores, while rare segment types—those with fewer training examples—exhibit lower and more variable F1 scores. This highlights the significant impact of data imbalance on classification performance. Additionally, the histogram reveals that most segment types have low F1 scores overall. These findings motivate the need for label merging, improved clustering, or data augmentation techniques to better handle rare segment types and improve overall classification quality.

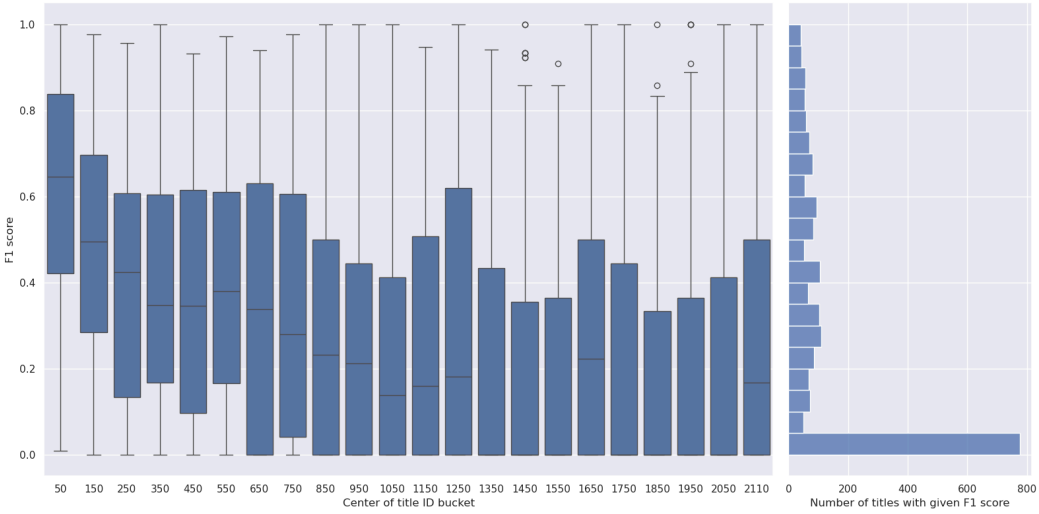


Figure 2: F1 scores for classification

We used both Topological Data Analysis (TDA) via the Mapper algorithm and Hierarchical Agglomerative Clustering to explore the structure of segment label embeddings. In the Mapper graph, tightly connected central regions indicate groups of segment labels that frequently co-occur or share similar semantic content(Figure 4). In contrast, peripheral or isolated nodes often represent rare or highly specific labels that show little overlap with others. To complement this, we applied hierarchical clustering and evaluated its quality using the silhouette score, which quantifies how well-separated the resulting clusters are. Based on the silhouette score trend, we selected k = 72 clusters to strike a balance between meaningful separation and preserving fine-grained structure among note labels. This combined approach helps organize the label space and supports future efforts to reduce label redundancy and improve interpretability.

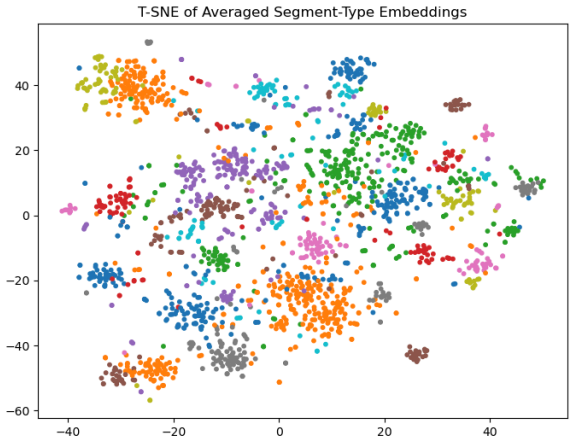


Figure 3: Hierarchical Clustering

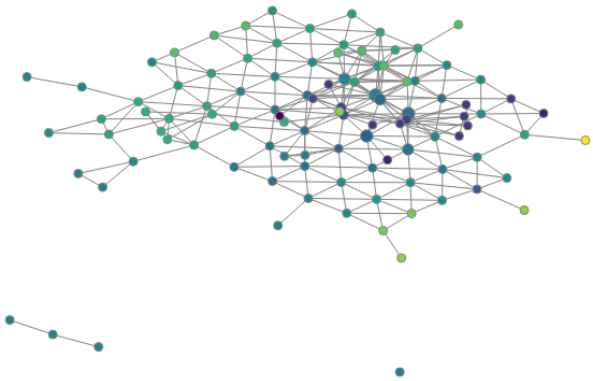


Figure 4: Mapper Clustering

Dataset: https://physionet.org/content/mimiciii/1.4/