Coherence-based Modeling of Clinical Concepts Inferred from Heterogeneous Clinical Notes for ICU Patient Risk Stratification

Tushaar Gangavarapu tusgan@amazon.com
Gokul S Krishnan gsk1692@gmail.com
Sowmya Kamath S sowmyakamath@nitk.edu.in

Healthcare Analytics and Language Engineering (HALE) Lab
Department of Information Technology, National Institute of Technology Karnataka, Surathkal, India

Modeling Patient-specific Information

The need for more patient-centric and precise assessments in the healthcare industry has motivated the development of intelligent Clinical Decision Support Systems (CDSSs) [1]. Advanced medical interventions in ICUs make patients vulnerable to several complications—the lack of accurate knowledge of the etiology of such complications leads to the inability to accurately stratify risk. Structured medical data in the form of Electronic Health Records (EHRs) contain numerical assessments (e.g., lab results) and are amenable to standard statistical analysis. However, unstructured clinical text and images also contain valuable information concerning the state of a patient. **Clinical nursing notes maintain objective and subjective assessments of a patient's condition**—can be utilized to uncover hidden clues about the mental state of a patient. **As these notes are informally written, modeling such notes is challenging due to their high-dimensionality, rawness, sparsity, complex linguistic and temporal nature, inconsistent abbreviations, and occurrence of rich medical jargon.** The voluminosity of nursing notes can be observed from the heavy-tailed distribution of the MIMIC-III nursing notes across various patients, with an average of 176.49 nursing notes per patient.

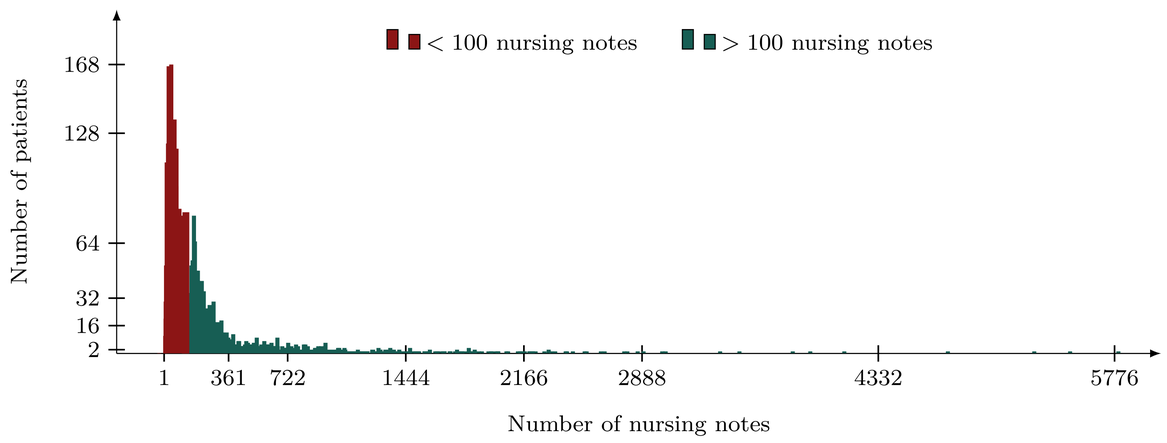

Figure 1: Distribution of the nursing notes across various MIMIC-III subjects.

Objectives

- Design of ***FarSight***, a long-term aggregation mechanism that employs future lookup to detect disease onset with the earliest recorded symptoms, to enable prioritized care and prevent further complications.
- Leveraging vector space and topic modeling approaches to derive optimal data representations from the unstructured clinical text, essential for accurate ICD-9 code group prediction. Our experimental results corroborate the efficacy of the proposed strategy when compared to state-of-the-art models built on structured patient data.
- Designing a technique that utilizes voluminous nursing notes for accurate risk stratification, thus eliminating the dependency on the availability of structured EHRs. This eliminates a significant roadblock in the development of CDSSs for hospitals in developing nations with low structured EHR adoption rates.

Coherence-based Modeling of Clinical Notes

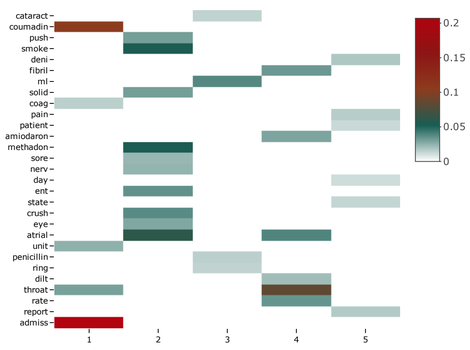

Figure 2: Correlations between top terms' membership in top five coherence-based LDA clusters.

Latent Dirichlet Allocation (LDA) is a cluster analysis approach based on the three-layer Bayesian framework: documents, topics, and tokens. LDA draws a mixture of topics from the Dirichlet distribution and facilitates a soft probabilistic clustering of tokens into topics and documents into topics. LDA posits that each term and clinical note belong to a set of clinical topics with a certain probability. Nonnegative Matrix Factorization (NMF) is a factorization approach that decomposes multivariate data into topics. In NMF, each topic is a nonnegative linear combination of the tokens in the vocabulary. NMF iteratively decomposes the data matrix ($N \times |\mathbb{V}|$) into two lower rank matrices with $\mathcal{T}$ topics ($N \times \mathcal{T}$ and $\mathcal{T} \times |\mathbb{V}|$). These topic models capture the context of occurrence and co-occurrence, which is essential for accurate predictability of the underlying deep neural models. Determining the optimal number of LDA or NMF clusters is a challenging task. To address this issue, we utilize the Topic Coherence (TC) [3] between the topics to derive the optimal number of clusters. Furthermore, when topics are learned from a multinomial distribution over words from noisy and sparse text data, they are less coherent and hard to interpret. **TC evaluates topic models with a greater guarantee of human interpretability.** In our work, we adopt LDA and NMF with TC (C-LDA and C-NMF) as TC accounts for the semantic similarity between the higher scoring tokens and facilitates the generation of human-understandable topics. Let $\mathcal{T} = \{t_1, t_2, \ldots, t_k\}$ be a topic generated from a topic model which is represented using its top-k most probable tokens (t_is). Note that higher values of the average pairwise similarity among the tokens in $\mathcal{T}$ imply greater coherence of the topic. For a predetermined similarity measure $S(t_i, t_j)$ (here NPMI), coherence score is computed as shown in (1):

$$\text{Coherence}_S(\mathcal{T}) = \frac{\sum_{1 \le i \le k-1} \sum_{i+1 \le j \le k} S(t_i, t_j)}{\binom{k}{2}} \tag{1}$$

where $t_i, t_j \in \mathcal{T}$. The coherence score comes from external data, i.e., the data not used during training, and is intended to regularize the topic models. The NPMI similarity score is an extension of the pointwise mutual information score, and is used in finding associations and collocations between the words.

$$\text{NPMI}(t_i, t_j) = \frac{\text{PMI}(t_i, t_j)}{-\log_2(\Pr(t_i, t_j))};\ \text{PMI}(t_i, t_j) = \log_2\left(\frac{\Pr(t_i, t_j)}{\Pr(t_i)\Pr(t_j)}\right) \tag{2}$$

The individual confirmation measures obtained for all topics ($\mathcal{T}_i$s) are averaged to obtain the final coherence score. The number of topics for both LDA and NMF models was determined to be 100, by computing the coherence score of several topic models.

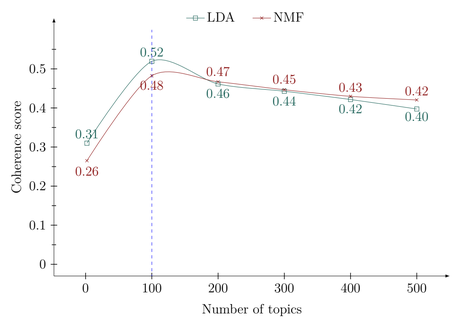

Figure 3: Coherence score comparison to determine the optimal number of topics.

Diagnostic ICD-9 Code Groups

ICD-9 codes are a taxonomy of diagnostic codes typically used by healthcare professionals. This study only focuses on group predictions, owing to the high granularity of the diagnostic codes—each code group comprises a set of similar diseases. **This study focuses on the risk stratification as a multi-label problem**, where each nursing note is mapped to multiple ICD-9 code groups. The ICD-9 codes for a given admission are mapped into 19 distinct code groups. The ICD-9 code range of $760 - 779$ was left out since it corresponds to the *conditions originating in the perinatal period*, which is usually assigned to newborns, who are excluded from this study. Additionally, to lower the computational cost of training, we merged all the reference and supplemental V-codes into a single code group. Although our work and the state-of-the-art [2] differ in data and cohort selection, both the works share similar statistics concerning the ICD-9 code groups, thus facilitating a fair comparison of performance.

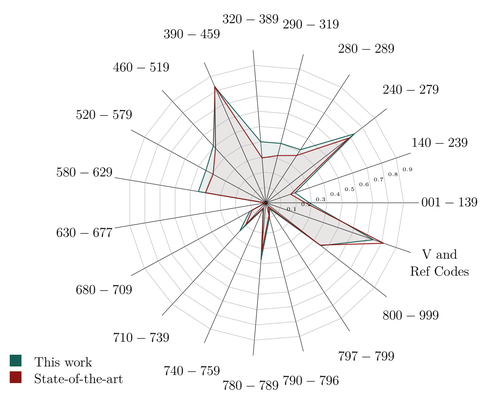

Figure 4: Comparison of ICD-9 code group statistics with the state-of-the-art model [2].

Acknowledgements

Funded by the Government of India's DST-SERB Early Career Research Grant (ECR/2017/001056) to Sowmya Kamath S.

Deep Neural Learning for Code Group Prediction

Two deep neural models, **Multi-layer Perceptron (MLP)** and **Attention LSTM (A-LSTM)** are employed for code group prediction—trained to minimize binary cross-entropy loss using an Adam optimizer (batch size of 128, eight epochs).

In MLP, the output of a neuron in every layer serves as an input to the subsequent layer. A neuron in the current layer (l) with the input $I^{(l)}$ is activated in the following layer ($l+1$) as $g^{(l)}(W^{(l)} \cdot I^{(l)} + b^{(l)})$, where $g^{(l)}$ is a non-linear activation such as Rectified Linear Unit (ReLU), tanh, or logistic sigmoid, and $b^{(l)}$ and $W^{(l)}$ are the bias and weight matrix at layer l. MLP uses backpropagation to determine the gradient of the loss function. This study employs an MLP network with one hidden layer of 75 nodes, activated using a ReLU function, and one output layer of 19 nodes, activated using a sigmoid function.

LSTM employs four gates, including the input gate i, the forget gate f, the output gate o, and the candidate value g for the cell state. The precise form of an LSTM update at a layer l and time step t is:

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} W^{(l)} \begin{pmatrix} h_{t-1}^{(l)} \\ h_t^{(l-1)} \end{pmatrix};\ c_t^{(l)} = f \odot c_{t-1}^{(l)} + i \odot g;\ h_t^{(l)} = o \odot \tanh(c_t^{(l)}) \tag{3}$$

We utilize the attention mechanism for the clinical task. Let H be the matrix of output vectors $[h_1, h_2, \ldots, h_T]$ produced from LSTM. The representation r_j of a nursing note η_j after T time steps is computed as $H \cdot (\text{softmax}(v^T \cdot \tanh(H)))^T$, where v is a trainable parameter. We use an attention LSTM with dimension size of 289 for the embedding (17 time steps) and 300 for the LSTM hidden state. The multi-label classification is facilitated using a sigmoid activation of the final A-LSTM output.

Unstructured Nursing Text Processing: Materials

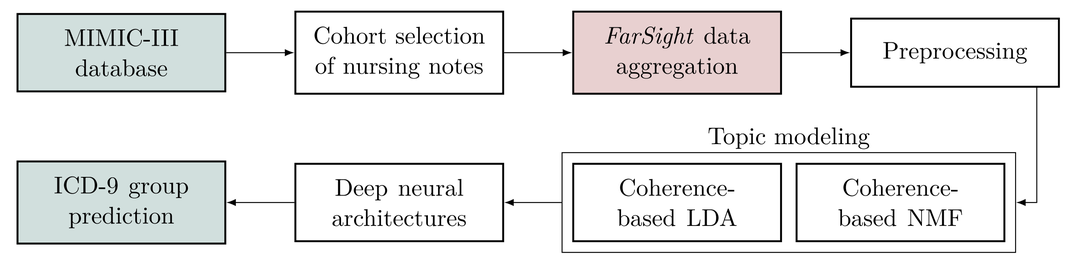

Figure 5: NLP pipeline used in the prediction of the ICD-9 code group.

Parameter	Total
Clinical nursing notes	223,556
Sentences in the nursing notes	5,244,541
Words in the nursing notes	79,988,065
Unique words in the nursing notes	715,821

Table 1: Statistics of the nursing note text corpus.

MIMIC-III provides comprehensive health data of over 40,000 ICU patients. Predefined selection criteria were employed for cohort selection—firstly, records corresponding to patients older than 15 were retained, and secondly, only the first hospital admission of a patient was considered. Erroneous entries were filtered out and duplicate patient records were identified and removed. The resultant dataset consisted of nursing notes corresponding to 6,532 patients, and **the data in these records were aggregated using the proposed *FarSight* technique**, which was designed to aggregate the patient data using a future lookup on all the detected diseases in the later medical records concerning that patient. If $\mathcal{P}$ is the set of all patients, and a patient p has a sequence of N clinical notes, then $\mathcal{S}^{(p)} = \{(\eta_i^{(p)}, \mathcal{I}_i^{(p)})\}_{i=1}^N$, with each clinical note $\eta_i^{(p)}$ is mapped to a code $\mathcal{I}_i^{(p)}$ indexed in chronological order. Now, *FarSight* aggregates the ICD-9 codes across the nursing notes of a patient using a future lookup, resulting in $\mathcal{S}^{(p)} = \{(\eta_i^{(p)}, \mathcal{I}^{(p)})\}_{i=1}^N$, where $\mathcal{I}^{(p)} = \{\mathcal{I}_i^{(p)}\}_{i=1}^N$. We aim at learning a function $\mathcal{F}$ to estimate the probability of classifying a given nursing note $\eta_j^{(p)}$ into a set of diagnostic code groups: $\mathcal{F}(S^{(p)}) \approx Pr(\mathcal{I}^{(p)} \mid \eta_j^p)$.

Results and Discussion

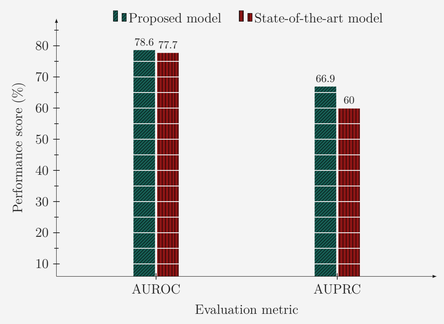

Figure 6: Comparison of the proposed approach with the state-of-the-art model [2].

AUPRC varies with changes in target class ratio, and hence is more informative than AUROC while evaluating imbalanced data. F1 score captures both precision and recall of the prediction, while, MCC accounts for true positives, false positives, and false negatives, thus serving as a balanced measure even with class imbalance. The existing works, including the state-of-the-art model [2], are built on the structured EHRs, modeled using numerical feature sets to aid in the prediction of clinical events. **Our model built on the unstructured medical text and preprocessed using the *FarSight* approach outperformed the state-of-the-art model by 11.50% in AUPRC and 1.16% in AUROC.** Furthermore, the existing works do not benchmark their performance only on AUPRC and AUROC metrics. *FarSight* effectively models the unstructured data to facilitate the detection of the onset of the disease with the earliest recorded symptoms, and such modeling results in an improvement in the clinical decision-making process. We observed that utilizing *FarSight* helps in accurate health risk appraisal well in advance, with an overall accuracy of 80%. Thus, CDSSs built on the predictive capabilities of *FarSight*-aggregated and C-LDA classified modeling could demonstrate effective patient-centric and evidence-based risk assessment, thus ensuring proper channeling of preventive and prioritized care.

Data Model	Classifier	Performance score				
		ACC	F1	MCC	AUPRC	AUROC
C-LDA (140,792 × 100)	MLP	**0.7954 ± 0.0003**	0.7175 ± 0.0008	**0.5743 ± 0.0006**	**0.6692 ± 0.0006**	**0.7857 ± 0.0004**
	A-LSTM	0.7932 ± 0.0002	**0.7186 ± 0.0002**	0.5712 ± 0.0007	0.6660 ± 0.0007	0.7854 ± 0.0013
C-NMF (140,792 × 100)	MLP	0.7826 ± 0.0004	0.7011 ± 0.0008	0.5480 ± 0.0007	0.6530 ± 0.0013	0.7735 ± 0.0006
	A-LSTM	0.7811 ± 0.0005	0.6990 ± 0.0040	0.5449 ± 0.0007	0.6510 ± 0.0009	0.7715 ± 0.0026
LDA (140,792 × 100)	MLP	0.7950 ± 0.0003	0.7168 ± 0.0020	0.5735 ± 0.0012	0.6685 ± 0.0013	0.7848 ± 0.0011
	A-LSTM	0.7930 ± 0.0007	0.7153 ± 0.0034	0.5701 ± 0.0022	0.6655 ± 0.0013	0.7833 ± 0.0020
NMF (140,792 × 100)	MLP	0.7829 ± 0.0006	0.7029 ± 0.0016	0.5498 ± 0.0009	0.6530 ± 0.0017	0.7744 ± 0.0007
	A-LSTM	0.7815 ± 0.0008	0.6935 ± 0.0052	0.5451 ± 0.0024	0.6535 ± 0.0014	0.7689 ± 0.0031

Table 2: Experimental results for ICD-9 code group prediction using MLP and A-LSTM.

Concluding Remarks

FarSight, a preventive care mechanism for detecting disease onset with earliest recorded symptoms is presented. Two coherence-based topic modeling approaches were employed to capture the semantic information in unstructured nursing notes and derive optimal representations with emphasis on human interpretability, further leveraged for ICD-9 code group prediction using deep neural models. We benchmarked the performance of the proposed models using several evaluation metrics essential in the accurate assessment of reliability and robustness. **The proposed model outperformed the structured EHR data based state-of-the-art model with an improvement of 11.50% in terms of AUPRC and 1.16% in terms of AUROC.** Moreover, our model eliminates the dependency on structured EHRs, typically a prerequisite requirement for the development of CDSSs, thus is extremely vital in countries with low EHR adoption rates.

References

[1] P. S. Mathew and A. S. Pillai. Big data solutions in healthcare: Problems and perspectives. In *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, pages 1–6. IEEE, March 2015.

[2] S. Purushotham, C. Meng, Z. Che, and Y. Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of Biomedical Informatics*, 83:112 – 134, 2018.

[3] M. Röder, A. Both, and A. Hinneburg. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408, New York, NY, USA, 2015. ACM.