Improving Mortality Prediction in ICU using Spark Big Data Tools

Vedanuj Goswami
Georgia Institute of Technology
vedanuj@gatech.edu

***Abstract*—Extracting topic models from patient notes and combining them with admission baseline feature can significantly improve Intensive Care Unit(ICU) mortality predictions. Using Apache Spark and its Machine Learning Pipeline this project builds a scalable model for mortality predictions using big data tools. The experiments are performed on the MIMIC III database. This project shows that using additional structured features like Simplified Acute Physiology Score (SAPS) I and II, Acute Physiology Score III (APSIII), Sepsis-related Organ Failure Assessment(SOFA) can improve baseline prediction with area under Receiver Operating Characteristic curve(AUROC) to 0.79. Using Logistic Regression and parameter exploration in ML Pipeline the combined time-varying AUROC score increased upto 0.94 for in hospital mortality. This significantly improves previous reported scores of baseline admission AUROC of 0.77 and combined time-varying AUROC of 0.855.**

***Index Terms*—Big data, ICU Mortality, Topic Modelling, Machine learning, Scalability**

MORTALITY prediction in Intensive Care Units (ICU) has been motivated primarily by the need to compare the efficacy of medications, care guidelines and surgeries. Prediction of patient mortality in ICU is a widely researched area due to the challenge faced in processing and analyzing huge amount of data required to come up with a good prediction model. Along with high accuracy of the model it should also be capable of giving predictions in a short span of time due to critical health care needs. Big Data technologies such as Apache Spark can help in real time processing, aggregation, predictions based on patient EMR data. Apache Spark is a fast and general engine for large-scale data processing and is 10x faster than Apache Hadoop framework[17].

The main objective of this project is to use Big Data tools to make a scalable predictive model and to improve Receiver Operating Characteristic(ROC) & Precision Recall(PR) AUC scores of mortality prediction by using baseline admission features and topic modelling on clinical notes.

I. Literature Survey

Predicting ICU mortality have seen numerous approaches and techniques. Most techniques have used structured data to build ICU mortality models. Such structured data have shown to improve baseline prediction. All these works have used the Receiver Operating Characteristic AUC as the measure for the performance of the models. Some of the works using structured features include Oxford Acute Severity of Illness Score(AUC 0.88) [8], APACHE (AUC 0.77) [9], SAPS II(0.77) [10] and SOFA(0.84) [16]. In addition to these, features like age, gender and other such attributes have also been used frequently. Although the structured features gave high AUC scores these were still not very accurate for using in practical conditions and hence efforts to improve the ICU mortality prediction has been an ongoing research area.

Ghassemi et al. [6] showed that time-varying models that combine latent topic features (derived using Latent Dirichlet Allocation) in clinical notes and baseline features improved in hospital mortality AUROC to 0.85. Topic modelling has been previously explored by Saria et al. [15] with F1 Score of 88.3, Ghassemi et al. [7] with retrospective AUROC 0.855 and Lehman et al. [11] who applied Hierarchical Dirichlet Process to achieve an AUROC of 0.82. This methods have resulted in significant improvement in the prediction and this paper focuses to improve it further.

Due to unbalanced classes in mortality prediction data sets it is important to observe the Precision Recall characteristics of the predicted classes. David & Goadrich[5] and Saito & Rehmsmeier [14] argues that for problems suffering from class imbalance, using an evaluation metric of Precision-Recall AUC is better than Receiver Operating Characteristic AUC. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate.

II. Approach and Implementation

A. Data and Pre-processing

In this project ICU data from the MIMIC III 1.4v database is used, which includes electronic medical records (EMRs) for 46520 ICU patients. Data is first loaded into PostgreSQL database and then queried using *psql*. Patients age, gender, SAPS scores, SAPS-II scores, SOFA scores, APS III scores are calculated. Patient mortality outcomes are calculated to determine 15759 in-hospital mortalities. The resultant data consists of patients who are older than 15 years and not organ donors.

For the above mentioned patients their clinical notes from nursing, physicians, labs, and radiology excluding discharge notes[6] are extracted. Documents are first tokenized to create the vocabularies and then removing stop-words from Onix word list[4]. The total vocabulary after these steps is 159026 for notes considered till 120 hour interval. The cohort consisted of notes for 45196 patients with 671,021 notes for the 120 hour interval. 70% of the patients were used for training and the rest 30% were used for test.

B. Structured Features

Structured features are considered among age, gender, SAPS, SOFA, SAPS II, APS III, OASIS scores. Minimum Redundancy Maximum Relevance (mRMR) algorithm is used for feature selection from these features. Maximum-relevance selection select features that correlate strongest to the classification variable and Minimum Redundancy select features not redundant with another feature. Finally for best classification the 6 features age, gender, SAPS, SOFA, SAPS II, APS III are used.

C. Topic Modelling

Topic Modelling is done using Latent Dirichlet Allocation(LDA) [3] with $k = 50$ topics with an Expectation-Maximization(EM) LDA Optimizer[2]. The hyper-parameters of the Dirichlet priors are set as the topic distributions $\alpha = 1 + \frac{50}{k}$ and the topic-word distributions $\beta = 1 + \frac{200}{numberWordsInVocab}$. The topic distributions are sampled with 100 iterations. This gave a 50 dimensional vector of topic probabilities for each note. For a particular time interval all the vectors for each patient are concatenated by taking the average.

The topic vectors are concatenated into a matrix $\mathcal{Q}$ where each element $\mathcal{Q}_{n,k}$ is the proportion of topic k in the note n. The probability of a topic to enrich in-hospital mortality can be calculated using the enrichment scores described by Martin[12] :

$$\theta_k = \frac{\sum_{n=1}^{N} \mathcal{Q}_{n,k} * y_n}{\sum_{n=1}^{N} \mathcal{Q}_{n,k}} \quad (1)$$

where y_n is the binary mortality outcome variable, which takes the values 0 when patient lives and 1 when patient dies.

D. Spark ML Pipeline

The system is implemented on Apache Spark as Spark's in-memory primitives provide enhanced performance[17]. Spark provides Structured Query Language(SQL) support and a distributed machine learning framework on top of Spark, MLlib, which has very high computational speed [13]. Spark is run on Amazon Web Services(AWS) Elastic Compute Cloud(EC2) clusters.

Spark ML provides a uniform set of high-level APIs that help users create and tune practical machine learning pipelines. Spark ML standardizes APIs for machine learning algorithms to make it easier to combine multiple algorithms into a single pipeline, or work flow. A parameter grid can be built over a pipeline object by specifying various sets of parameters across different stages of the pipeline. The various stages of the pipeline can be specified in the form of a Directed Acyclic Graph (DAG). The input feature files are in SVMlib format. The ML pipeline is shown in Fig 1.

A Logistic Regression[1] is instantiated using Spark ML pipeline with ParamGridBuilder containing the regression parameters and maximum iterations. The LR estimator is then cross validated(10-fold) with a CrossValidator that takes the estimator, a Binary Classification evaluator and a ParamGridBuilder object. The best model is generated to determine the optimal values with AUC as an objective and is then used to make predictions on the test set.

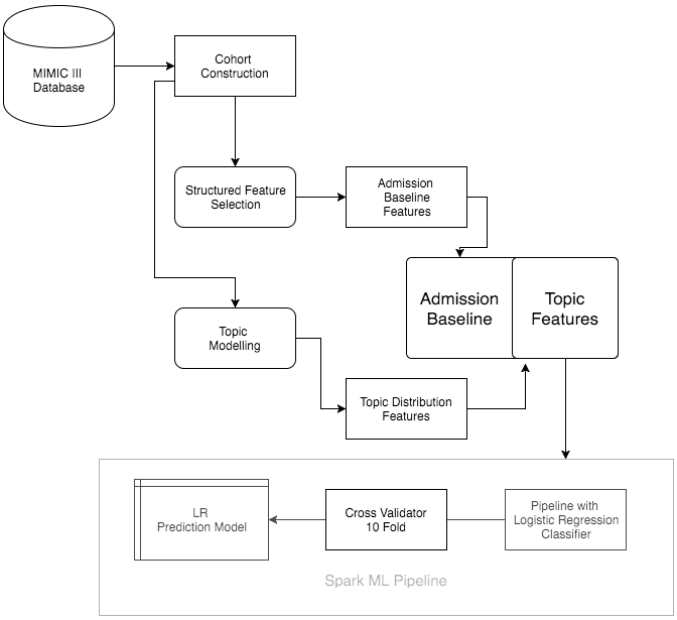

Fig. 1. Spark ML Pipeline with Cross Validator for least error model.

E. Prediction

Two prediction regimes are used : Admission baseline and Dynamic (time-varying) outcome prediction for in-hospital mortality.

1) Admission Baseline Model: : Only the structured features age, gender, SAPS II, SOFA, APS III and SAPS are used. Total of 6 features.

2) Time-varying Combined Model: : In addition to the structured features used in the admission baseline, time varying topic model features are used for this. The time window chosen is 12 hours. The notes are taken in a incremental stepwise fashion for 10 intervals. The notes for the patients who are discharged or expired in previous intervals were not discarded since those notes also can contain important information for training the LR. To be specific "discharge" note category is not included as this mention the mortality condition of the patient. These topic features (a total of 50) are added to the baseline features to make a total of 56 features for prediction.

F. Scalability

To improve the run-time and promptness of building the predictive model varying master-slave cluster combinations are used. Increasing the cluster size can improve the speed and run-time significantly.

III. Experiment Design and Evaluation

A. Topic Enrichment

LDA algorithm was run for 100 iterations and 50 topics are generated. The distributions of the in-hospital mortality

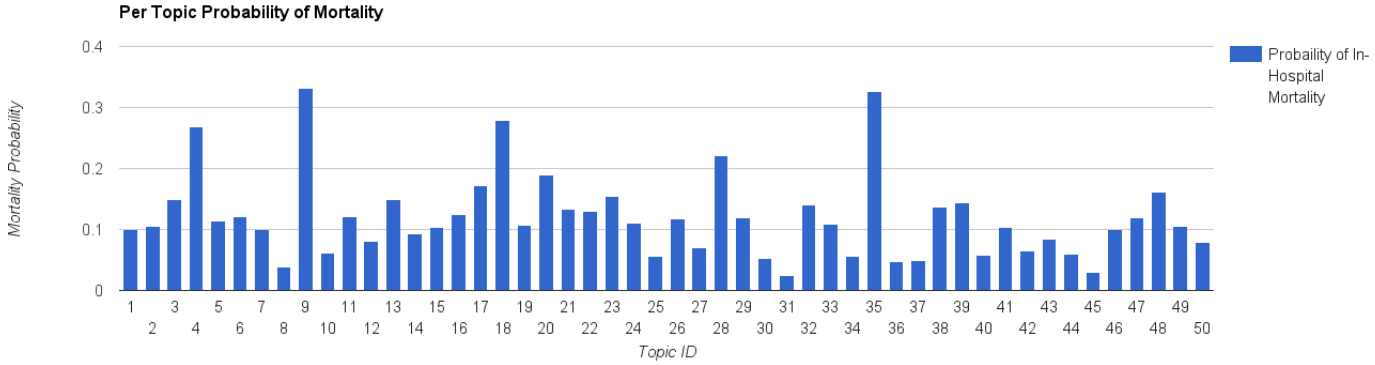

Fig. 2. The probability of in-hospital mortality for each topic, indicating that topics represent differences in outcome. Probabilities are calculated using Equation 1. Each bar shows the prevalence of a given topic k in the mortality category, as compared to the set of all patients.

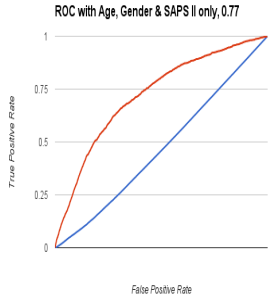

Fig. 3. ROC Curve with baseline features age, gender and SAPS II only.

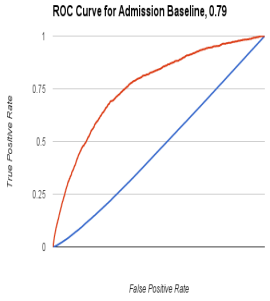

Fig. 4. ROC Curve with baseline features age, gender, SAPS II, SAPS, SOFA and APS III.

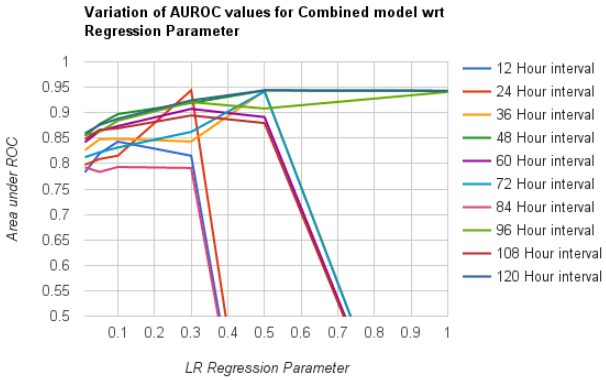

Fig. 5. Variation of AUC values with change in Regression parameter. Considering a single regression parameter for all time varying intervals can lead to reduced predictive scores. For some time intervals AUROC peaks at regression parameter 0.3 while for some others it peaks at 0.5. This shows a need to explore various possibilities of regression parameters for best model performance.

for each topics are shown in Fig 2. The topics showing high probability presents a good view of the possible causes for death in ICU. Topic 35 which talks about care, comfort family has the highest value as this may indicate end of life care options. Other topics which shows high probability for in hospital mortality are Topic 9 (cardiac arrest), Topic 18(acute renal), Topic 48 (ventilation respiratory) which are among highest mortality causes.

B. Predictive analysis

To improve upon the predicted scores in [6] various experiments are done.

1) Improving Admission Baseline prediction: Using only age, gender and SAPS II scores feature set an AUC of 0.77 was achieved at ICU admission previously. Severity score like SAPS, SOFA and APS III are added to these three features to check improvement in the baseline prediction AUC. These set of features were chosen using an mRMR algorithm which best improves the accuracy. With the inclusion of the additional features, the AUROC increases to 0.795 as can be observed in Fig 3 and Fig 4. This is significant improvement over the baseline model prediction scores in [6].

2) Improving Combined Time-varying model prediction: To improve the Time-varying model combined with Admission baseline features some new methods are tried.

- Instead of dropping the patients from the learning set who were discharged or expired before any interval, those patients' data are kept in the training model for all time intervals. However this data does not include any discharge notes. Only the notes prior to discharge are kept even for the patients who are discharged or who expired. This is because discharge notes specifically mention the expire status of a patient. The hypothesis behind this whole approach is that specific notes before the death or discharge of a patient(excluding "discharge" notes) can be helpful for topic modelling even though they did not complete a time interval in the time-varying model.
- To improve the predictive model the Spark ML pipeline is utilised to maximize the model performance. Logistic Regression parameters are set after a wide range of grid searching over the regression parameters and maximum number of iterations to find the best set of parameters. The best model is found using a 10 fold cross validation along with a Binary Classifier evaluator. The parameter grid search is done because the AUROC values vary a lot based upon the set of parameters chosen. Fig 5 shows how

regression parameter for the LR changes the AUROC. Hence selecting any one value for *regression parameter* without searching for the entire range of values can lead to lower AUC values.

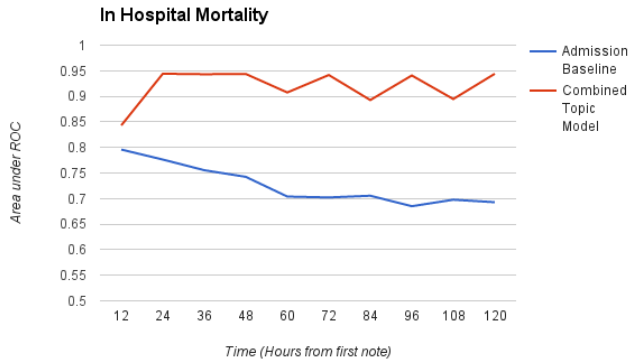

Fig. 6. LR model performance measured via area under ROC curve for in-hospital mortality. AUROC is higher for combined time-varying model. At 120 hour interval the score is 0.94

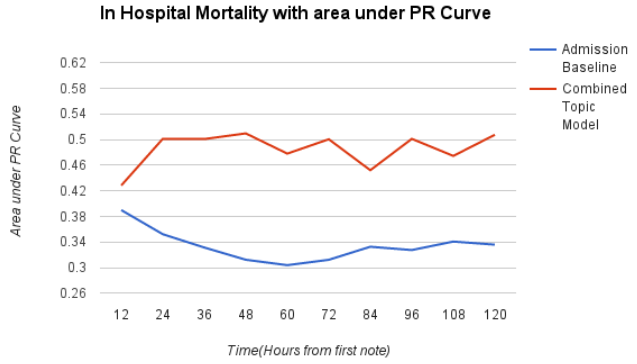

Fig. 7. LR model performance measured via area under PR curve for in-hospital mortality.

Augmenting the Topic modelling with these new methods improved the results(Fig 6). The AUROC increased to as high as 0.94 for 120 hour interval in-hospital mortality. This is improvement over 0.85 AUC prediction in the outline paper[6]. From Fig 6 we also observe that with increasing time the test patient number decreases as fewer patients have long ICU stays. This reduces the Baseline prediction scores significantly. However the time-varying combined model continues to perform well as observed in [6]. The above claim is also substantiated from the observation made from Fig 7 where the area under PR curve is much lower for admission baseline as compared to the combined time varying model. Hence even if AUROC as a best measure for unbalanced class data sets is debatable, the claim that topic model improves prediction significantly is verified by the AUPRC curves too.

3) Performance of Logistic Regression Classifier vs Support Vector Machine Classifier: Using Logistic Regression classifier in the ML Pipeline gave very good results for the AUROC values. Same experiments are performed using a Support Vector Machine Classifier with Stochastic Gradient Descent(SGD). The SVM with SGD was run by varying the number of iterations from 10 to 500 and the best performance model was chosen using 10 fold cross validation. The results are then compared with that of LR.

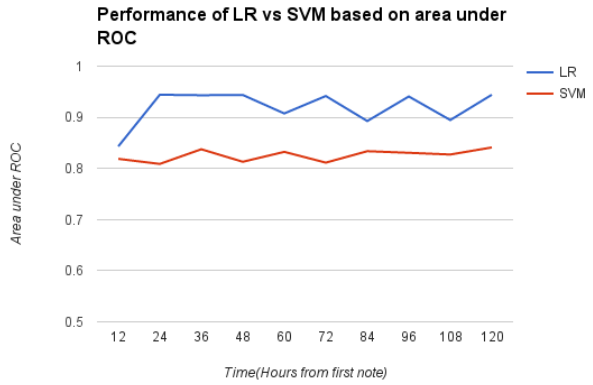

Fig. 8. Performance of LR vs SVM measured with area under ROC curve on the Combined Topic model predictions.

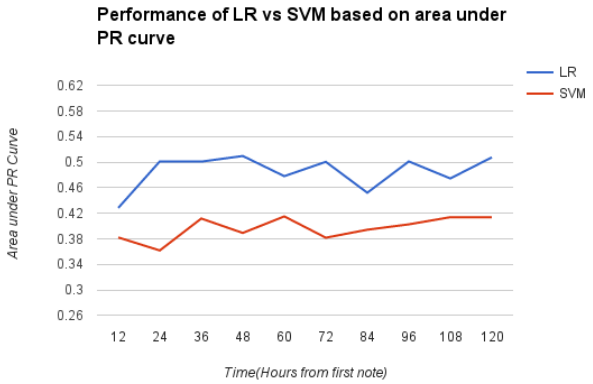

Fig. 9. Performance of LR vs SVM measured with area under PR curve on the Combined Topic model predictions.

Form Fig 8 and Fig 9 it is observed that the LR outperforms SVM on this data set both based on ROC and PR characteristics.

C. Improving Performance using Cluster Size increase

Topic modelling, feature construction and prediction over the large MIMIC III data set is very time consuming with a single node. Increasing the number of cluster nodes showed improvement in running times and faster performance. Fig 6 shows how fast LDA modelling runs for varying master and slave cluster combinations (m4xlarge EC2 instances).

IV. Discussions and Conclusion

Ghassemi et. al.[6] discovered that there is rich and useful information in patient notes and time-varying topic models can potentially be useful in predicting ICU mortality accurately.

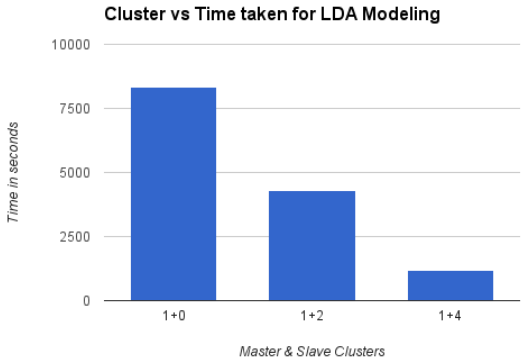

Fig. 10. LDA model building time vs number of Spark clusters. With 1 master and 0 slave clusters run time for 100 iteration LDA is 8345 seconds. Increasing the slave clusters to 4 reduces this time to below 1500 seconds.

Due to huge amount of medical data needed to be processed for use in real world, this project has proposed how big data tools(Spark) can be used to make a scalable model based on topic modelling for mortality predictions.

It is observed that some topics contribute more towards mortality and hence are crucial for building predictive models. The topics that show high probabilities often refer to situations which are very severe and often lead to expiry of the patient.

This project also showed significant increase in the in-hospital mortality AUC that can be attributed towards wide range of parameter exploration made possible by the Spark ML Pipeline with a Logistic Regression(LR) classifier and also due to the retaining of patient information in the training model. Another explanation is that the MIMIC III data set consists of twice the number of patient data in comparison to MIMIC II over which the experiments in [6] were performed. This significantly increased the number of training points and hence better model could be created.

The overall AUC score pattern(both ROC and PR) over the varying time intervals are similar to what observed in [6]. Combining baseline features with time varying topic models greatly increase the AUC scores for long term patients even though admission baseline prediction scores gradually decrease with time.

Logistic Regression classifier helps to provide a better explanation to the difference in patient features over a short interval of time since it is based on probabilities directly related to the features. This is in contrast to Support Vector Machines(SVM) which provide probabilities based on the distance from the separating hyperplane.

In order to improve the predictive capability of the model better topic models can be created with more number of LDA iterations and choosing only a threshold number of terms for each patient that are most predictive/relevant. In order to improve the scalability of the systems experiments with more number of clusters can be performed. More work needs to be done to determine better combination of topic models and parameters for building more robust and accurate predictive model for ICU mortality.

ACKNOWLEDGEMENT

This project and paper is completed under the guidance of Prof. Jimeng Sun. I would also like to thank and acknowledge the whole teaching team for their prompt and ready help/responses to our queries and doubts over the course of the project.

REFERENCES

[1] C. M. Bishop. Pattern recognition. *Machine Learning*, 2006.
[2] D. Blei and J. Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.
[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.
[4] C. Buckley and G. Salton. Stop word list. *UR L http://www. lextek. com/manuals/onix/stopwords2. html*, 2013.
[5] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.
[6] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits. Unfolding Physiological State: Mortality Modelling in Intensive Care Units. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 75–84, New York, NY, USA, 2014. ACM.
[7] M. Ghassemi, T. Naumann, R. Joshi, and A. Rumshisky. Topic models for mortality modeling in intensive care units. In *ICML Machine Learning for Clinical Data Analysis Workshop*, 2012.
[8] A. E. Johnson, A. A. Kramer, and G. D. Clifford. A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy*. *Critical care medicine*, 41(7):1711–1718, 2013.
[9] W. A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, and A. Damiano. The apache iii prognostic system. risk prediction of hospital mortality for critically ill hospitalized adults. *Chest Journal*, 100(6):1619–1636, 1991.
[10] J.-R. Le Gall, S. Lemeshow, and F. Saulnier. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24):2957–2963, 1993.
[11] L.-W. H. Lehman, M. Saeed, W. J. Long, J. Lee, and R. G. Mark. Risk stratification of icu patients using topic models inferred from unstructured progress notes. In *AMIA*. Citeseer, 2012.
[12] B. M. Marlin, D. C. Kale, R. G. Khemani, and R. C. Wetzel. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 389–398. ACM, 2012.
[13] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, et al. Mllib: Machine learning in apache spark. *arXiv preprint arXiv:1505.06807*, 2015.
[14] T. Saito and M. Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.
[15] S. Saria, G. McElvain, A. K. Rajani, A. A. Penn, and D. L. Koller. Combining structured and free-text data for automatic coding of patient outcomes. In *AMIA Annu Symp Proc*, volume 2012, pages 712–716. Citeseer, 2010.
[16] J.-L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. Reinhart, P. Suter, and L. Thijs. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive care medicine*, 22(7):707–710, 1996.
[17] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: Cluster computing with working sets. *HotCloud*, 10:10–10, 2010.

APPENDIX

A. Supplement Material

URL for video presentation :
https://youtu.be/KcCMpd6SGkw

B. Top 10 words for all the 50 topics

Topic Number	Top 10 terms
1	patient aspiration swallow speech status mental unable left weakness thin
2	total review patient code radial respiratory assessment admission family medical
3	left tube chest line catheter final clip report underlying post
4	vent peep remains fentanyl increased wean sedated versed cont insulin
5	line picc catheter placement left identifier procedure wire final report
6	abdominal bowel pain surgery patient abdomen surgical fluid colon obstruction
7	artery left carotid identifier aneurysm internal common angio vertebral evidence
8	pain denies oriented taking sleep diet floor chair alert sats
9	arrest cardiac dopamine patient family neuro unresponsive continue found monitor
10	patient total history code review continue medical signs radial dose
11	lasix sats resp increased mask monitor cont crackles pulmonary status
12	patient history setting hold recent daily hypotension home total baseline
13	chest reason underlying report tube admitting left final medical interval
14	wife neuro oriented nursing unable note hand speaking times family
15	ercp pancreatitis pancreatic patient abdominal biliary acute fluid stent transferred
16	head left contrast hemorrhage frontal report intracranial subdural admitting mass
17	mental altered status continue patient acute head respiratory blood noted
18	renal levophed acute line hypotension respiratory fluid pressors continue lumen
19	pleural effusion pericardial chest left drain fluid effusions echo thoracentesis
20	liver lactulose cirrhosis patient fluid continue total hepatic paracentesis ascites
21	contrast abdomen pelvis evidence report clip final left admitting free
22	intubated propofol tube respiratory sedated fentanyl sedation extubated endotracheal extubation
23	mass cancer lung metastatic cell patient left breast tumor biopsy
24	lasix heart systolic continue aortic acute ventricular pulmonary failure valve
25	etoh alcohol ciwa ativan withdrawal haldol valium agitated abuse agitation
26	renal dialysis kidney chronic esrd stage acute failure line catheter
27	fracture left spine fall cervical trauma spinal multiple fractures injury
28	thick vent remains suctioned yellow care resp cont coarse eyes
29	respiratory pulmonary patient copd continue acute home chronic sats bipap
30	pain left sicu sodium insulin tube dilaudid heparin total sliding
31	pain wean weaned insulin cabg chest lungs pacer wires csru
32	skin wound care impaired left integrity continue dressing foot ulcer
33	blood fever cultures culture vancomycin continue urine fluid sepsis temp
34	insulin blood type diabetes sliding patient continue glucose scale dose
35	family woman husband care daughter female support morphine team comfort
36	chest left aortic tube reason pleural report post underlying admitting
37	pain total chronic code morphine control respiratory patient transferred heart
38	liver hepatic portal transplant vein normal flow gallbladder report admitting
39	nursing micu note stool remains progress received sats foley urine
40	chest cardiac cath patient heparin pain continue coronary plavix daily
41	afib atrial rate fibrillation amiodarone coumadin continue patient diltiazem metoprolol
42	foley monitor urine yellow draining pain pulses skin adequate sounds
43	bleeding blood bleed units stable unit prbc active gastrointestinal upper
44	seizure patient psych history found seizures status total medical level
45	insulin graft bypass artery temporary start coronary chest patient tube
46	trach respiratory tube patient rehab picc tracheostomy bronch stent tracheal
47	neuro left head checks hemorrhage dilantin patient total exam insulin
48	lung assessment ventilation breathing tube respiratory ideal airway total sputum
49	heparin left lower pulmonary extremity bilateral femoral filter venous veins
50	cath groin iabp cont site note pulses remains progress follow