Improving Hospital Mortality Prediction with Medical Named Entities and Multimodal Learning

**Mengqi Jin, Mohammad Taha Bahadori, Aaron Colak,
Parminder Bhatia, Busra Celikkaya, Ram Bhakta, Selvan Senthivel, Mohammed Khalilia,
Daniel Navarro, Borui Zhang, Tiberiu Doman, Arun Ravi, Matthieu Liger, Taha Kass-hout**

Amazon.com Services Inc.
{maggjin, bahadorm, parmib, busrac, hitenram, ssenthiv, khallia,
navarda, bzha, tddoman, ravarun, ligerm, tahak}@amazon.com
aaron.r.colak@gmail.com

Abstract

Clinical text provides essential information to estimate the acuity of a patient during hospital stays in addition to structured clinical data. In this study, we explore how clinical text can complement a clinical predictive learning task. We leverage an internal medical natural language processing service to perform named entity extraction and negation detection on clinical notes and compose selected entities into a new text corpus to train document representations. We then propose a multimodal neural network to jointly train time series signals and unstructured clinical text representations to predict the in-hospital mortality risk for ICU patients. Our model outperforms the benchmark by 2% AUC.

1 Introduction

A growing number of studies in the healthcare domain have shown compelling results by applying deep neural networks on predictive modeling tasks where traditional methods have met bottlenecks. Recurrent Neural Networks (RNN) and their variants such as Long Short-Term Memory (LSTM) were some of the earliest deep neural networks to analyze real-valued measurements [Lipton et al., 2015, Che et al., 2018] and higher dimensional structured claims data [Choi et al., 2016]. Convolutional architectures also have been found to be faster and achieve similar accuracy [Razavian et al., 2016, Liu et al., 2018]. In both cases, the learning objective was to stratify patients based on their risks of encountering certain clinical events, such as mortality and disease onset.

Recently there has been increasing interest in utilizing unstructured information from clinical notes to improve clinical events prediction performance, as free-text notes paint a more elaborate picture of the patient. Topic modeling is commonly used to extract insights from notes [Ghassemi et al., 2015, Miotto et al., 2016, Rumshisky et al., 2016, Suresh et al., 2017]. Boag et al. [2018] compared notes' representations generated by Bag of Words (BoW), Word2Vec and LSTM by evaluating their performance on downstream clinical prediction tasks. Their results showed no simple winning algorithm could ensure the best performance across all tasks. BoW perfomed well on tasks where outcomes were strongly correlated with certain phrases or words; LSTM, on the other hand, performed better on tasks for which temporal information was important. For in-hospital mortality prediction, BoW and Word2Vec achieved similar performance and outperformed LSTM. In this study, we choose a Doc2VecC algorithm [Chen, 2017] to represent notes because of its proven success in capturing semantic information [Lau and Baldwin, 2016] (details on text representation are in Section 2.2).

Meanwhile, results from using proprietary healthcare data are not easily reproducible, resulting in difficulties in making comparisons between different studies and algorithms. Outside of making

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

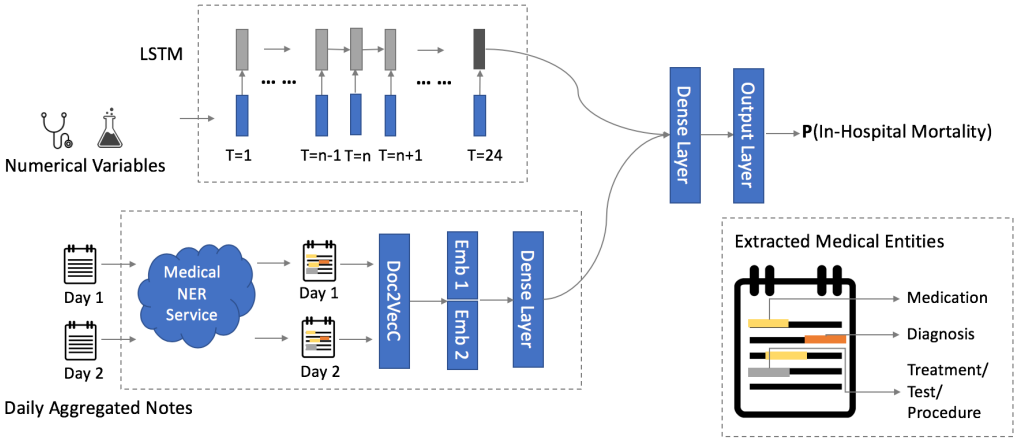


Figure 1: Data processing and multimodal architecture of proposed model: selected named entities are extracted from daily aggregated notes to train document embeddings to pass through a shallow feedforward neural network before jointly trained with last output from LSTM to make predictions.

data public and accessible, additional efforts are expected to ensure repeatable experiments, such as the well described study design and publicly available cohort construction code [Johnson et al., 2017]. Sharing the same goals, we selected a published in-hospital mortality prediction benchmark [Harutyunyan et al., 2017] which uses only structured data (explained in Section 3). We adopt the same cohort construction pipeline, and continuously iterate upon their model to ensure comparable and reproducible results.

Our work is distinguishable from other studies in three aspects: 1) We leverage a medical natural language processing service for named entity recognition and negation detection to process text beyond traditional text wrangling techniques; 2) We experiment with a multimodal model architecture to incorporate notes' information; 3) We also quantify the model's performance gain realized from the clinical notes (which are directly comparable to the published benchmark).

2 Data extraction and text representation

We conduct experiments on MIMIC-III data set [Johnson et al., 2016]. We adopt the same data extraction pipeline as the benchmark [Harutyunyan et al., 2017] to prepare baseline features and introduce note assignment events. The data analysis flowchart of the best-performed model is pictured in Figure 1. Overall, there are 42,276 ICU stays extracted from 33,798 patients at least 18 years old at admission. First 48 hrs data of ICU admission are collected to predict whether the patient encounters a fatality throughout her stay. The same split of 70%/15%/15% and seed state are used to create train/validation/test sets.

2.1 Baseline feature set

The benchmark cohort contains 17 signals derived from the structured database including vital signs and lab results. In the fixed 48 hrs observation window, signals are discretized every 2 hours, which yields 24 signals per stay. Missing values are imputed from the previous record, and population statistics if the previous record was unavailable. This feature set is referred to as **Vital**.

2.2 Clinical text representation

We first examine the completeness of timestamps associated with the note to propose model structures. ECG and Echo reports do not have time stamps, but only a date available, which take up 10.49% of the first two days' non-discharge notes. In light of the incompleteness of date-only time-stamped notes, we concatenated clinical notes assigned in the first two days of admission into two aggregated notes. Discharge notes are excluded from the text corpus to avoid information leak. We replaced numbers

in notes with zeros and low-frequency words that have appeared less than 10 times with a special token. A Continuous Bag-Of-Words based document representation algorithm, Document Vector through Corruption (Doc2VecC) [Chen, 2017] is selected to learn note embeddings. In this approach, the original document is corrupted by randomly removing a significant portion of words, and the document is represented using only the embeddings of the remaining words. We select Doc2VecC because it offers speedup during training as it significantly reduces the number of parameters to update in backpropagation. Another advantage of this method is that weighing down frequent words across document corpora can yield performance improvements.

Embeddings of tokenized notes We use an internal medical text tokenization service to process daily aggregated notes. We then represent tokenized notes using Doc2VecC algorithm after following the preprocessing steps introduced above and obtain a vocabulary size of 108,907. This set of notes' representations is referred to as **NoteEmb**.

Embeddings of concatenated relevant medical entities Clinical text can be full of redundant information, therefore requires proper pre-processing such as named entity recognition (NER), and negation scope detection to filter negated information [Miotto and Weng, 2015]. Contrary to previous studies [Miotto et al., 2016, Liu et al., 2018] that tag clinical narratives using medical ontology look-up and regular expressions to detect negations, such as NegEx [Chapman et al., 2001], in this paper, we leverage a neural network based internal medical NER service [Shen et al., 2017]. The NER system has a hierarchical architecture composed of three components: 1) a convolutional character-level encoder extracting features for each word from characters, 2) a convolutional word-level encoder extracting features from the surrounding sequence of words, and 3) an LSTM tag decoder inducing a probability distribution over any sequences of tags. We extract five types of entities including medical condition, medication, tests, treatments, and procedures. The medical NER service also jointly identifies negated entities and returns a negation tag as an attribute. We then exclude negated entities from the text corpus before training embeddings. For example, a phrase "no oropharyngeal lesion" has "oropharyngeal lesion" recognized as a medical condition type of entity with negation attribute. This entity is discarded from the text corpus as it indicates an absence of the condition. The vocabulary size of extracted entities corpus is 25,503. Embeddings trained on top of concatenated named entities is referred to as **EntityEmb**.

3 Learning

We experiment with two neural network architectures to model the medical text. We first apply a LSTM on concatenated baseline features with pre-trained text embeddings. Daily concatenated notes are assigned timestamps until the end of the day. Missing embeddings are imputed in the same way as baseline features. Second, we test a multi-modal network to process structured signals and embeddings separately, where baseline features are passed through LSTM, and two pre-trained daily embedding vectors are concatenated to pass through a feedforward network and are jointly trained afterward.

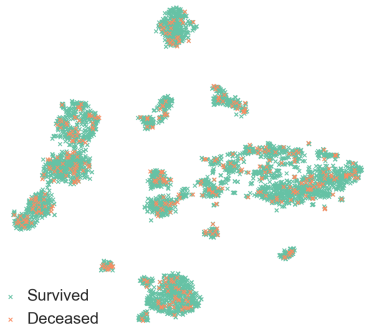
Benchmark model architecture The benchmark model is a LSTM neural network and uses only structured data. Given a series of evenly spaced signals within a fixed observation window, $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$, where $\mathbf{x}_t \in \mathbf{R}^L$ and L is the number of input signals, the model learns a series of hidden state vectors $\mathbf{h}_1, ..., \mathbf{h}_T = \text{LSTM}(\mathbf{x}_1, ..., \mathbf{x}_T)$, and uses the last hidden vector $\mathbf{h}_T$ to predict outcome label y.

Proposed multimodal neural network Multimodal deep learning frameworks were introduced by [Ngiam et al., 2011]. When multiple modalities are available, the fused representation can be used as input for discriminative tasks [Srivastava and Salakhutdinov, 2012]. We concatenate two days' text embeddings $\mathbf{e}_1$ and $\mathbf{e}_2$, and pass it through one layer feedforward network to learn a hidden representation $\mathbf{h}_e$. Next, we concatenate $\mathbf{h}_e$ and last output from LSTM $\mathbf{h}_T$ to learn a joint representation $\mathbf{h}_j$ of clinical variables and text embeddings through a feedforward network to predict class label $y \in 0, 1$.

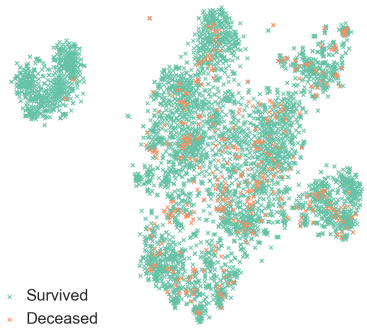
4 Experiments

4.1 Experimental Setup

We establish the selected benchmark with the same parameter settings (**Vital**) [Harutyunyan et al., 2017]. We then compare mortality prediction performance of our two neural network structures



(a) aggregated raw notes embeddings.



(b) aggregated entity embeddings.

Figure 2: t-SNE plots for embeddings, color-coded by associated outcomes.

Table 1: Experimental Results

Model			
Feature Set	Neural Network Structure	AU-ROC (%)	AU-PRC (%)
Vital (Benchmark)	LSTM	0.8531 ±0.0020	0.5030 ±0.0051
Vital + NoteEmb	LSTM	0.8496 ±0.0018	0.5040 ±0.0050
Vital + NoteEmb	Multi-modal	0.8669 ±0.0018	0.5310 ±0.0051
Vital + EntityEmb	LSTM	0.8703 ±0.0017	0.5470 ±0.0048
Vital + EntityEmb	Multi-modal	**0.8734 ±0.0019**	**0.5290 ±0.0056**

(LSTM and Multimodal) using two feature sets (**Vital + NoteEmb** and **Vital + EntityEmb**). No dropout nor normalization is applied. The hidden layer size of LSTM is set to 256. The number of hidden units of the dense layers for text representation and joint representation are set to 100 and 300 respectively. Parameters are cross-validated on the validation set. Learning rate is set to 0.0001. Our experiments are implemented with Lasagne version 0.2.1 [Dieleman et al., 2015], and run on Amazon Web Services' p2. 8xlarge GPU instances.

4.2 t-SNE plot of medical text embeddings

We apply the t-SNE algorithm [van der Maaten and Hinton, 2008] to visualize embeddings generated by the Doc2VecC algorithm. We show two t-SNE plots for notes' embeddings (Figure 2a) and aggregated named entities' embeddings (Figure 2b). Embedding vectors are colored based on their associated clinical outcomes. It's interesting to see that the notes embeddings and entities embeddings are projected into different patterns. Moreover, a cluster of entities embeddings is mostly associated with negative outcomes. Whereas, the positive and negative cases can be observed in most subgroups of notes embeddings. This suggests embeddings generated from entities correlates with the final outcome of an ICU admission.

4.3 Prediction results

We train each model 20 times with different initialization seeds. The best-performed model on the validation set is selected by the F1 score, and bootstrapped on the test set over 100 re-sampling. Our results show the proposed multimodal neural network is more efficient to incorporate text information with structured data comparing to LSTM. Meanwhile, models using embeddings trained from medical named entities present consistently better performance than those trained from the simple tokenized text. Our results are directly comparable and reproducible by adopting the same data processing flow as the benchmark study. In the future, we would like to further apply entity linking techniques to normalize extracted entities, and explore different algorithms to represent notes information to enhance predictive modeling of clinical events.

References

Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318, 2016.

Narges Razavian, Jake Marcus, and David Sontag. Multi-task prediction of disease onsets from longitudinal laboratory tests. In *Machine Learning for Healthcare Conference*, pages 73–100, 2016.

Jingshu Liu, Zachariah Zhang, and Narges Razavian. Deep ehr: Chronic disease prediction using medical notes. *arXiv preprint arXiv:1808.04928*, 2018.

Marzyeh Ghassemi, Marco AF Pimentel, Tristan Naumann, Thomas Brennan, David A Clifton, Peter Szolovits, and Mengling Feng. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data. In *AAAI*, pages 446–453, 2015.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6:26094, 2016.

A Rumshisky, M Ghassemi, T Naumann, P Szolovits, VM Castro, TH McCoy, and RH Perlis. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational psychiatry*, 6(10):e921, 2016.

Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding using deep networks. *arXiv preprint arXiv:1705.08498*, 2017.

Willie Boag, Dustin Doss, Tristan Naumann, and Peter Szolovits. What's in a note? unpacking predictive value in clinical note representations. *AMIA Summits on Translational Science Proceedings*, 2017:26, 2018.

Minmin Chen. Efficient vector representation for documents through corruption. *arXiv preprint arXiv:1707.02377*, 2017.

Jey Han Lau and Timothy Baldwin. An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*, 2016.

Alistair EW Johnson, Tom J Pollard, and Roger G Mark. Reproducibility in critical care: a mortality prediction case study. In *Machine Learning for Healthcare Conference*, pages 361–376, 2017.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771*, 2017.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

Riccardo Miotto and Chunhua Weng. Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials. *Journal of the American Medical Informatics Association*, 22(e1):e141–e150, 2015.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.

Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*, 2017.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.

Sander Dieleman, Jan Schlüter, Colin Raffel, Eben Olson, Søren Kaae Sønderby, Daniel Nouri, Daniel Maturana, Martin Thoma, Eric Battenberg, Jack Kelly, et al. Lasagne: first release. *Zenodo: Geneva, Switzerland*, 3, 2015.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.