

Explainable Clinical Decision Support from Text

Jinyue Feng

Unity Health Toronto, Toronto ON
jinyue@cs.toronto.edu

Chantal Shaib

Unity Health Toronto, Toronto ON
chantal.shaib@unityhealth.to

Frank Rudzicz

Unity Health Toronto, Toronto ON
Department of Computer Science, University of Toronto, Toronto ON
Vector Institute for Artificial Intelligence, Toronto ON
Surgical Safety Technologies, Toronto ON
frank@cs.toronto.edu

Abstract

Clinical prediction models often use structured variables and provide outcomes that are not readily interpretable by clinicians. Further, free-text medical notes may contain information not immediately available in structured variables. We propose a hierarchical CNN-transformer model with explicit attention as an interpretable, multi-task clinical language model, which achieves an AUROC of 0.75 and 0.78 on sepsis and mortality prediction on the English MIMIC-III dataset, respectively. We also explore the relationships between learned features from structured and unstructured variables using projection-weighted canonical correlation analysis. Finally, we outline a protocol to evaluate model usability in a clinical decision support context. From domain-expert evaluations, our model generates informative rationales that have promising real-life applications.

1 Introduction

Electronic medical records (EMRs) store both structured data (e.g., vitals and laboratory measurements) and unstructured data (e.g., nursing and physician notes). Previous clinical prediction tasks have focused on structured data (e.g., [Desautels et al., 2016](#); [Gultepe et al., 2013](#); [Ghassemi et al., 2014](#)) which, despite their utility, may not capture all of the useful information in associated text. Clinical decision support systems rarely take advantage of free-text notes due to the complex nature of clinical language and interpretation. Rules and specialized grammars can be applied to circumvent issues around clinical language; however, these methods rely on the presence of certain phrases and spelling, and do not account for the highly variable note structures across departments and hospitals ([Yao et al., 2019](#); [Mykowiecka et al., 2009](#); [Assale et al., 2019](#)). Further, opaque models without explainability are often met with resistance in medical

contexts ([Challen et al., 2019](#); [Ahmad et al., 2018](#); [Gordon et al., 2019](#)). To address these challenges, we propose a novel multi-task language model that also provides rationales for decisions in medicine.

Our multi-task model leverages ClinicalBERT ([Alsentzer et al., 2019](#)), which is a transformer-based model pre-trained on clinical corpora. Given the uniqueness of medical text, we introduce a combination of CNN and transformer encoders to capture phrase-level patterns and global contextual relationships. Additionally, we explore latent attention layers to generate rationales.

Based on availability, we use the MIMIC-III database ([Johnson et al., 2016](#)) to predict two outcomes: sepsis and mortality in the intensive care unit (ICU). All experiments are conducted on notes written in English. We define the task of sepsis prediction more rigorously than previous work due both to using textual data only, and to emphasize the practicality of this model in real-world applications. Moreover, we use canonical correlation analysis (CCA; [Hotelling 1992](#)) to explore relationships between latent features learned from both structured and unstructured data. Finally, we propose an evaluation protocol to examine the usability of our model as an interpretable decision support tool.

2 Related work

2.1 Transformers in the clinical domain

Transformers ([Vaswani et al., 2017](#)) have gained popularity given their strong performance and parallelizability. The success of the transformer-based BERT ([Devlin et al., 2019](#)) has inspired numerous studies to apply it in various domains. For example, BioBERT was pretrained on PubMed abstracts and articles and was able to better identify biomedical entities and boundaries than base BERT ([Lee et al., 2020](#)). [Alsentzer et al. \(2019\)](#) further fine-tuned

BioBERT on the MIMIC-III clinical dataset (Johnson et al., 2016) and released the model as ClinicalBERT. We use these pretrained BERT-based models as static feature extractors and build layers upon the word embeddings to learn task-specific representations spanning long documents.

2.2 Language model explainability

Explainable AI is an emerging field with no standardized methodology or evaluation metrics. The definition of model explainability also varies by application; however, a generally accepted approach to language model explainability is through extractive rationales (Lei et al., 2016; Mullenbach et al., 2018; Wiegrefe and Pinter, 2019).

The wide application of attention mechanisms has led to an ongoing debate over whether attention can be used as explanation (Serrano and Smith, 2019; Jain and Wallace, 2019; Wiegrefe and Pinter, 2019). Jain and Wallace (2019) claimed that attention scores in recurrent neural networks (RNNs) did not correlate with other feature-importance measures, and adversarial attentions did not affect model predictions, concluding that attention was not explanation. Wiegrefe and Pinter (2019) challenged these assumptions by proposing diagnostic tests that allow for meaningful interpretation of attention, but also showed that adversarial attention distributions failed to achieve the same level of prediction performance as real model attention.

We propose a clinical decision support tool that uses explanations to enhance model usability and reliability. Therefore, we adopt a view similar to that of Wiegrefe and Pinter (2019), in that attention provides plausible rationales for use in practice, even though it may not provide a complete internal representation of the model’s behaviour (Serrano and Smith, 2019; Jain and Wallace, 2019).

2.3 Clinical tasks

Sepsis is an extreme systemic inflammatory response to infection. If left untreated, sepsis can lead to life-threatening complications such as organ failure and septic shock. The ability to predict sepsis before symptom onset allows for earlier intervention, thus improving patient outcomes. Previous work on sepsis detection focused on both post-hoc identification as well as predicting the need for early intervention from structured data (Desautels et al., 2016; Taylor et al., 2016; Nemati et al., 2018; Gultepe et al., 2013). As mortality has an explicit label in EMRs, the focus has been

on expiry likelihood for early intervention rather than post-hoc identification (Ghassemi et al., 2014; Grnarova et al., 2016). We focus on work that used the MIMIC-III database (Johnson et al., 2016).

Insight (Desautels et al., 2016) provided a method for predicting sepsis from vital signs within a fixed-time window before suspected onset on retrospective data. Gultepe et al. (2013) proposed a similar structured-data model for mortality and sepsis prediction; however, the features were pre-selected and only considered five measurements. While these methods achieved robust results compared to traditional clinical measures (e.g., MEWS, qSOFA, SIRS; Churpek et al. 2017), none took advantage of the unstructured data found in EMRs.

Culliton et al. (2017) claimed that unstructured data in EMRs contain information not found in the structured variables. They used GloVe word embeddings to represent notes for each patient, and only excluded discharge summaries to minimize explicit mentions of sepsis. Simply excluding discharge summaries, however, is not sufficient to avoid label leakage – a diagnosis may appear in the notes as the clinician becomes aware of symptoms. We carefully filter notes to ensure no label leakage occurs and further refine our definition of sepsis prediction, as described in Section 4. Ghassemi et al. (2014) used topic modeling for textual representations aggregated with structured patient data to predict mortality, but Grnarova et al. (2016) showed that using convolutional document embeddings for each patient outperformed these topic modelling strategies for mortality prediction. Similarly, we deploy convolutional layers in our model to obtain sentence-level embeddings. Horng et al. combined structured and unstructured data for sepsis prediction, using topic models and continuous-bag-of-words (CBOW) to represent text. Despite success, GloVe word embeddings, topic models, and CBOW do not generally capture the complexity and contextual relationships between words in a given text. Specifically, these methods rely primarily on word frequency and collapse multiple meanings of a word into a single representation. To this end, we implement a transformer-based model to represent our clinical notes, which we hypothesize may capture the contextual complexity between tokens more completely.

3 Methods

3.1 Model architectures

The structure of our model is illustrated in Figure 1. We now explain each component in detail.

BERT word embeddings: BERT and its variants have exhibited strong performance in various tasks and we are interested in its application specifically in medical contexts. As shown in Figure 2, medical documents can easily contain thousands of tokens. With the sequence length limit of 512 tokens, using BERT as a fine-tuning language model on long documents is practically challenging or impossible. Instead, we approach this problem in a depth-first manner and use BERT as a static feature extractor on a sentence-by-sentence basis. Such a feature-based approach with BERT has proved to be nearly as effective as the fine-tuning approach in other tasks (Devlin et al., 2019).

We split each document into n sentences of m tokens and use a separate data loader with a sequential sampler to group them into sub-batches. The input is truncated or padded at both the sentence- and token-level. We then feed the sentences into a BERT model and take the mean of the last four encoder layers as token embeddings. For tokenization, we omit two irrelevant tokens [CLS], which is used as a pooling mechanism in fine-tuning models, and [SEP], which is used in next sentence prediction and sentence-pair classification tasks. BERT-related modeling and processing code comes from HuggingFace’s Transformers library (Wolf et al., 2019).

Given an input $T = [t_{11}, t_{12} \dots t_{ij} \dots t_{nm}]$, where t_{ij} denotes the j^{th} token of the i^{th} sentence, the BERT feature extractor outputs

$$X = [x_{11} \dots x_{nm}] = BERT(T),$$

where x_{ij} is a d_{emb} -dimensional vector (i.e., the hidden dimension of the BERT configuration) corresponding to t_{ij} .

Convolutional layer: Previous studies using CNNs to process medical notes have achieved good results on tasks such as mortality prediction and ICD-9-CM diagnosis code classification (Grnarova et al., 2016; Mullenbach et al., 2018; Si and Roberts, 2019). Specifically, a qualitative evaluation of text snippets from an attentional CNN indicated the model’s ability to learn features that are deemed informative and diagnosis-relevant by a physician (Mullenbach et al., 2018). This suggests

that the CNN is suitable for extracting information regarding patient status at the phrase-level. We use a simple 1D convolutional layer along the sequence of each sentence followed by ReLU activation and 1D max-pooling to obtain sentence representations.

Taking X as the input, the CNN outputs an $n \times d_{feature}$ matrix.

$$S = MaxPool(ReLU(Conv(X)))$$

where $d_{feature}$ is the number of output channels of the convolution layer.

Transformer patient encoder: Medical notes frequently contain repeated segments of medical histories as well as plans for future treatment. Although related work in patient-clinician dialogue has explicitly used time-series information (Khattak et al., 2019), the strict temporal order of patient conditions in clinical notes can be disrupted by repeating information. Yet, the highly complex mechanisms of medical outcomes entail that the co-existence of some conditions may change the indication of others. We apply a two-layer transformer encoder on top of sentence features to capture a unified representation among descriptions. This step of encoding results in a matrix

$$S_T = Transformer(S)$$

that shares the same dimension as S .

Although multi-head attention is powerful (Clark et al., 2019), it is not yet clear how to derive rationales for model prediction from such an approach. For model explainability, we instead apply an explicit attention mechanism that is directly implementable and interpretable.

Latent attention: The outputs of the transformer encoders are sentence-level features. To obtain patient representations, we use a latent attention mechanism adapted from similar work in if-then program synthesis (Liu et al., 2016). The goal of latent attention is to dedicate a component of the model to explicitly learning the importance of each unit of explanation such as the sentence or word.

The latent attention scores are calculated from sentence features using a position-wise feed-forward network (Vaswani et al., 2017). Given S_T , an n -dimensional vector a_{input} is computed as

$$a_{input} = FeedForward(S_T)$$

and the attention weight is

$$a = Softmax(a_{input} + a_{mask}),$$

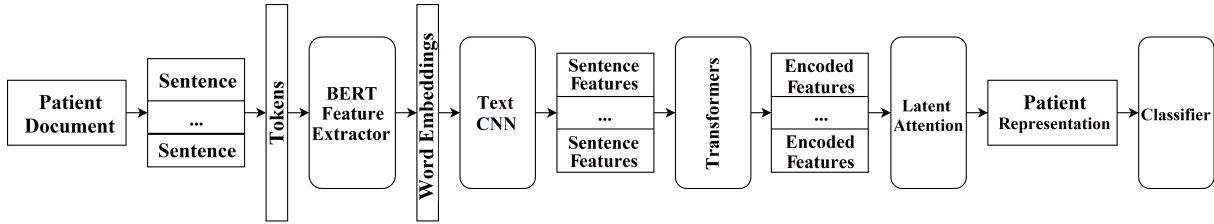


Figure 1: Model architecture and data flow. Each patient document undergoes various levels of feature extraction to arrive at token-, sentence-, and patient-level representations. The explicit attention layer provides a latent representation for a patient. The final, attended-to patient representation is used in the classification task.

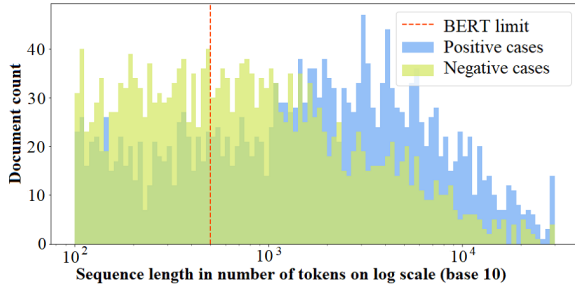


Figure 2: Distribution of documents based on token lengths for the mortality dataset. 2842 out of 5147 documents exceed the token limits of BERT, indicated by the vertical dashed line.

where a_{mask} is an n -dimensional vector for which values unmasked positions are 0 and values at padding positions are $-10,000$.

The final $n_{feature}$ -dimensional patient vector p is computed as the weighted sum of sentence features, which we can define as the dot product,

$$p = \sum_{i=1}^n S_{T_i} a_i = S_T \cdot a$$

and feeds a linear layer and a softmax classifier.

3.2 Canonical Correlation Analysis

Classic canonical correlation analysis (CCA) provides a set of linear transformations that maximally correlate data points from multiple views (Hotelling, 1992). We use projection-weighted CCA (PWCCA) (Morcos et al., 2018) to investigate the correlation between learned textual features and various structured data that are split into their respective clinical tests, shown in Table 1. Given two vectors, $\mathbf{x} \in \mathbb{R}^{d \times n}$ and $\mathbf{y} \in \mathbb{R}^{d \times m}$, where n and m denote feature dimensions and d denotes number of data points, the objective is

$$(w_1^*, w_2^*) = \arg \max_{w_1, w_2} \frac{w_1' K_{XY} w_2}{\sqrt{w_1' K_{XX} w_1 w_2' K_{YY} w_2}},$$

Clinical test	Related structured variable
Complete Blood Count (CBC)	Hemoglobin Hematocrit; Mean Corpuscular Hemoglobin; Platelets; Red Blood Cell Count; White Blood Cell Count
Prothrombin Time (PT)	Partial Thromboplastin Time; Prothrombin Time Inr; Prothrombin Time Pt
Urea, Creatinine, and Electrolytes (UCE)	Bicarbonate; Blood Urea Nitrogen; Chloride; Creatinine; Potassium; Sodium
Arterial Blood Gases (ABG)	Anion Gap; CO ₂ (etco ₂ , pco ₂ , etc.); Partial Pressure of Carbon Dioxide; pH
Blood Pressure (BP)	Central Venous Pressure; Diastolic Blood Pressure; Mean Blood Pressure; Pulmonary Artery Pressure Systolic; Systolic Blood Pressure
Individual Tests (IND)	Glucose; Calcium; Calcium Ionized; Magnesium; Phosphate; Phosphorous; Glasgow Coma Scale Total
Pulmonary Flowmetry (PF)	Fraction Inspired Oxygen Set; Peak Inspiratory Pressure; Positive End-Expiratory Pressure Set; Respiratory Rate; Tidal Volume Observed
Primary Vitals (PV)	Heart Rate; Oxygen Saturation; Temperature

Table 1: Mapping of clinical tests to their corresponding structured variables.

where K_{XY} denotes the cross covariance and K_{XX} and K_{YY} denote the covariances.

Following the method of singular value CCA (Raghu et al., 2017), we use singular value decomposition to obtain the weights w_1, w_2 . From this, we get a total of $\min\{n, m\}$ canonical correlation coefficients. The high dimensionality of the feature representations may result in noisy coefficients that hinder the similarity measurements. We use projection weighting to compute a weighted mean of the canonical variates, which accounts for the importance of CCA vectors relative to the original input (Morcos et al., 2018). The PWCCA similarity between vectors \mathbf{x} and \mathbf{y} is computed with

$$d_{pwcca}(x, y) = 1 - \sum_c \alpha_i \rho^{(i)}$$

where α_i denotes the normalized importance weights, and $\rho^{(i)}$ the i^{th} CCA coefficient. We use an open-source implementation of PWCCA¹ in our experiments. Understanding the correlated information in patient features between textual and structured data may provide insight on what latent information is learnt from the text.

¹<https://github.com/google/svcca/>

4 Data

MIMIC-III: MIMIC-III is a clinical database comprising de-identified EMRs of 58,976 hospital admissions to the critical care units of the Beth Israel Deaconess Medical Center (Johnson et al., 2016). All variables are recorded between 2001 and 2012. Note that, although ClinicalBERT is pretrained on MIMIC-III, this does not preclude its use from downstream tasks on the same dataset; Alsentzer et al. emphasize that any impact is negligible given the size of the entire MIMIC-III corpus compared to sub-sampled task corpora. In this study, we choose sepsis and mortality tasks because these are the standard tasks of this dataset. However, our model is not specifically tailored to these tasks, and may be generalized to wide range of potential applications.

Data preprocessing: To avoid data leakage among hospital admissions of the same patient, we only include patients with one hospital admission. We select adult patients from the single-admission group and obtain a base population of 31,245 hospital admissions. We randomly sample negative cases to balance the dataset in both tasks.

For text, we concatenate text from different note entries into one document for each patient and remove punctuation (except periods and commas), masked identifiers, digits, and single characters. When merging patients' notes, we remove sentences that have already appeared in previous notes to avoid repetition. The notes are appended in chronological order according to their timestamps and truncated to a maximum of 50,000 tokens.

For mortality prediction, we do not differentiate note types. For sepsis, we find differences in the frequencies of note *types* between positive and negative populations, which may result in a trivially learned solution. After consulting with clinicians, we exclude note types that are irrelevant to sepsis and select nursing and physician notes only.

Whereas structured variables have explicit timestamps that can be easily related to symptom onset, the timestamp of a note may not. For example, a note containing descriptions of possible infection may be entered after antibiotic administration. Anchoring notes with lab measurement timestamps significantly limits the number of positive cases in our dataset, especially when compared to other studies containing similar sepsis cohorts (Section 2.3). Nonetheless, we view the imposed time-

window constraints as necessary to create an honest representation of *prediction*. Discharge summaries and any notes written after patient outcomes occurred are excluded to avoid direct access to the solution. Unfortunately, these steps are not always taken in the literature.

For the structured data used in Section 3.2, we use MIMIC-Extract² to ensure a standard patient population. After obtaining time-binned cohort data, we extract measurements within the same time frames as the selected notes.

Sepsis: Systemic inflammatory response syndrome (SIRS), characterized by abnormal body temperature, heart rate, respiratory rate, and white blood cell count, often precedes sepsis. In this task, we aim to predict whether a patient in SIRS would become septic. In contrast to previous work where the negative sepsis populations did not necessarily have SIRS (Section 2.3), our task is more restrictive, as the model must learn features that are distinctive of sepsis onset rather than general indications of SIRS. We use ICD-9-CM codes to label cases, where patients with codes for explicit sepsis, or a combination of infection and either organ failure or SIRS, are considered positive. Although ICD-9-CM codes can be unreliable (O'Malley et al., 2005), we use multiple criteria to deal with false negatives and SIRS as a filter to avoid false positives (Angus and Wax, 2001). We notice that very few notes are recorded before the first onset of SIRS, possibly due to a time delay in writing or logging notes. To compensate for the lack of data, notes before and within 24 hours of the first onset of SIRS are included. To avoid possible label leakage, we remove sentences containing mentions of “*sepsis*” or “*septic*”. The final cohort contains 1262 positive cases and 1500 negative cases.

In-ICU mortality: MIMIC-III has an expiry timestamp for patients who died in the hospital, which identifies the positive cohort for in-ICU mortality prediction. To ensure that all samples represent patient conditions in the ICU, we only include notes written within ICU stays. The dataset has 2562 positive cases and 2587 negative cases.

²https://github.com/MLforHealth/MIMIC_Extract

5 Experiments

Our experiments explore 1) differences in prediction due to pretraining, 2) multiview projection, and 3) *evaluable* explainable AI.

5.1 Clinical vs Non-Clinical BERT.

To compare the effect of pretraining BERT with domain-specific clinical data on the overall quality and performance of the model, we substitute BioBERT (Lee et al., 2020) and base BERT (Devlin et al., 2019) as the token embedding component. We run both sepsis and mortality tasks on the different *BERT models and compare the final performance. The results are shown in Table 2.

In comparing performance between tasks, the models achieve better performance in mortality than sepsis. Considering that patients in the negative cases in sepsis task all had SIRS, which is one of the diagnostic criteria of sepsis, the high false positive rate among all three models is expected.

ClinicalBERT models converge faster and outperform the other two models in both sepsis and mortality tasks. BioBERT and BERT models are comparable in performance; however, BioBERT models exhibit a tendency to output positive results, resulting in high recall and high false positive rates. The fact that BioBERT does not perform better than base BERT suggests that clinical-specific pretraining is crucial and cannot be replaced by pretraining on general biomedical corpora.

5.2 Structured vs Textual Data

To investigate the relationships between patient features extracted from structured and text data, we separately train RNN models to learn representations from different groups (see Table 1) of laboratory measurements, and we conduct PWCCA (Figure 3) to compute their similarities to patient features from the language model.

Structured data model: To obtain a single vector from time-series structured data, we construct a 2-layer single-directional GRU network followed by a linear layer to project the mean GRU output to a feature vector that has the same dimension as the language model feature vectors. Only the patients that appear in the language model cohort are selected. Each model is trained for 50 epochs, and the best-performing one is used to extract features.

CCA details: To avoid spurious correlations typically found in small datasets, the number of data

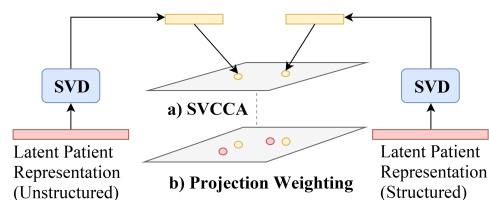


Figure 3: Visualization of PWCCA. The patient representations are taken from the models before the classifier. First, a) a latent space is learned with SVCCA; then, b) The original representation is projected onto the learned latent space, and the PWCCA is computed.

points (n_{sample}) should be at least five times³ the feature dimension ($d_{feature}$). Therefore, we include all shared patients between structured and unstructured datasets, and over-sample the data for the sepsis task. We set up random baselines for each test where we randomly generate n_{sample} $d_{feature}$ -dimensional vectors using the same sampling strategy as the real features. To ensure that our features are meaningful, we only analyze features extracted by models that reach an AUROC of at least 0.75. It is important to note that we constructed the structured dataset to obtain the patient representation, not to compare model performance. The structured inputs contain measurements after the onset of patient outcomes, so the metrics should not be compared to those of the language model. Additionally, the structured data models fail to learn to predict sepsis from SIRS cohort, so we include negative samples without SIRS whose data are extracted from random time frames. Model performance and PWCCA similarity (described by Morcos et al. (2018)) are listed in Table 3.

Feature correlation: The similarity scores are subject to confounding factors such as noise and sample size. Due to limited data availability, we can only comment on the general patterns. The structured data model and language model converge to correlated solutions, compared to random baselines. We do not observe any clear relationship between structured model performance and similarity. The features learned from *all* lab measurements, which supposedly encode a more comprehensive patient representation than any subgroup alone, are close to the features learned from medical notes, especially in the mortality task. For the sepsis task, the test groups that are highly related

³Experiments demonstrating the choice of sample sizes in CCA can be found at <https://github.com/googlesvcca>

Model	Sepsis				Mortality			
	AUROC	F1	Precision	Recall	AUROC	F1	Precision	Recall
BERT	0.72	69.3	64.3	75.0	0.75	74.2	77.7	70.9
BioBERT	0.72	71.2	59.8	88.1	0.76	76.8	72.6	81.6
ClinicalBERT	0.75	73.0	64.4	84.3	0.78	78.9	78.2	79.7

Table 2: Test performance scores using different BERT models.

Features	Sepsis		Mortality	
	AUROC	Similarity	AUROC	Similarity
All	0.75	0.68	0.92	0.762
CBC	0.77	0.80	0.5	-
PT	0.76	0.60	0.5	-
UCE	0.68	-	0.57	-
ABG	0.77	0.60	0.62	-
BP	0.76	0.65	0.5	-
IND	0.77	0.93	0.88	0.686
PF	0.78	0.61	0.62	-
PV	0.5	-	0.5	-
Random	-	0.45	-	0.361

Table 3: Structured model test performance and PWCCA similarity to text features. The *All* category encompasses all test groups and their features. Table 1 shows the full list of features and their corresponding test categories.

to systematic inflammation or organ dysfunction (CBC, BP, IND) show especially strong correlation with the textual features. The results suggest that our language models learn to encode the most relevant patient conditions for each outcome. Future work includes further examining representation correlations, and other multi-view models combining structured and unstructured data as inputs.

5.3 Evaluating Explanations

Evaluating model explainability remains a broad area of research. Our primary objective is a usable model that can be deployed as a real-life decision support tool. Therefore, we focus on human evaluation as our assessment of rationale quality. We outline a novel evaluation protocol that measures the quality of the extracted rationales by leveraging clinical domain expertise. To avoid arbitrary judgments, we work with the physician to tailor the definition of utility for each task; this is expanded upon in the Appendix along with a stand-alone quantitative evaluation on non-clinical data of latent attention as an explanation mechanism.

To obtain succinct meaningful explanations, we calculate an attention threshold score

$$a_{threshold} = \max\left(\frac{1}{n_s}, a_{sentence_i}\right),$$

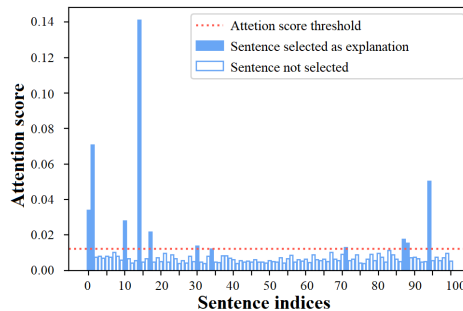


Figure 4: Example attention distribution over sentences in one patient document.

where a denotes attention scores, n_s is the number of sentences, and $i = \min(20, \lceil \frac{n_s}{10} \rceil)$. This ensures that selected sentences have higher attention scores than uniform attention and at most 10% of the original texts are included. To avoid burdening the evaluator, at most 20 sentences are selected for documents with more than 200 sentences. Figure 4 shows an example distribution of attention scores and demonstrates our explanation generation criteria. To prevent overly complicated results, we only evaluate the correctly predicted cases.

All independent evaluation uses a command-line user interface.

5.3.1 Labeling task

Labeling is designed to evaluate the informativeness of our generated explanations. Sentences are presented sequentially to an expert physician who chooses at each step to either predict patient outcome or check the next sentence. Sepsis has defined diagnosis criteria that must be followed in clinical practice, and information about such criteria are not necessarily available even in complete documents. However, mortality risk assessment, despite its difficulty, is common in critical care. Therefore, we only conduct the labeling task on the mortality dataset. We compare human predictions to those of our model and note the number of selected sentences necessary for each prediction. A test case fails if the evaluator does not make a decision after reviewing all selected sentences. This

	Pos	Neg	Total
N_{cases}	119	136	255
Conclusion	98.4%	98.5%	98.5%
Correctness	69.2%	96.3%	82.7%
Sentences Read (c)	4.0	3.5	-
Sentences Read (i)	4.2	8.2	-

Table 4: Labeling task results. We list the number of cases, percentage of concluded cases out of all cases, percentage of correct cases out of total concluded cases, and the average number of sentences read for both correct (c) and incorrect (i) cases.

method evaluates whether the attended sentences are sufficient to provide enough information for a clinical decision, and empirically evaluates the number of sentences needed for rationales.

The results are presented in Table 4. On average, the evaluator reaches a correct conclusion in mortality prediction 82.7% of the time by reading approximately 4 sentences per case (or a selected 0.5% of the note, on average). Such evidence strongly suggests that our model is capable of extracting the most relevant information from long documents. We also observe a general pattern that fewer sentences are needed for a correctly predicted case, which indicates that the ordering of sentences based on attention is generally reliable.

Interestingly, the evaluator almost correctly predicts all negative cases but not positive cases in the mortality task. Multiple reasons may account for the high false negative rate. First, mortality prediction is an intrinsically challenging task for humans. A bias towards survival may naturally occur when a sentence can be interpreted differently based on various contexts. Second, explanations for negative cases are more likely to be independent from the contextual information that are not included in the rationales. Our evaluator comments that a seemingly poor patient condition may translate to completely opposite outcomes depending on the coexistence of other conditions. In real-life applications, providing full documents with highlighted explanations may be an easy solution that helps to direct users’ attention to the most important parts without losing reference to additional contexts.

5.3.2 Rating task

In a second evaluation, we sample cases not used in the labeling task. We present model predictions and the entirety of the rationales sentence-by-sentence to an expert physician. The physician is instructed

	Sepsis			Mortality		
	Pos	Neg	Total	Pos	Neg	Total
$N_{sentences}$	1016	464	1480	958	486	1444
N_{cases}	64	54	118	76	52	128
$\%_{helpful, All}$	41.8	95.0	-	61.7	82.7	72.2
$\%_{helpful, Top 4}$	-	-	-	75.9	86.4	80.0
$\%_{helpful, Cases}$	96.4	74.1	86.0	-	-	-

Table 5: Rating task results.

to decide whether each sentence in the rationale contains information that helps explain the model decision. To avoid arbitrary judgements, we work with the physician to develop clear definitions of explanation utility, as shown in the appendix. This method assesses the average informativeness of selected sentences as well as the usability of our model for the purpose of clinical decision support.

Given the characteristics of mortality and sepsis (see the appendix for a detailed discussion), the evaluation is meaningful at the sentence- and case-levels for the two tasks. Table 5 summarizes the results. Between the positive and negatives cases, an average of 72.2% of sentences in the mortality task and 86% of cases in the sepsis task are rated as helpful for understanding model decisions. A closer look at the results shows that 80% of the first four sentences are rated as helpful, which indicates that the specific algorithm that generates rationales should be refined in future work to further exclude sentences with lower attention scores (see Figure 4). Nonetheless, the application of our model as an explainable decision support tool is very promising.

6 Conclusion

Language can provide valuable support to improve clinical decision-making. We conduct a diverse set of experiments to explore several aspects of the applicability of deep NLP in the clinical domain. We also address challenges in extracting medical documents that are representative of a predictive task.

We augment the power of domain-specific BERT and build a hierarchical CNN-Transformer that can potentially be applied to any long-document processing task. The model achieves AUROC scores of 0.75 and 0.78 on sepsis and mortality tasks, respectively. We also address model explainability by experimenting with a simple (yet effective) linear attention mechanism, and emphasize the interaction between models and users in the design of a novel protocol to evaluate explanations. Not only

are we able to sufficiently predict cases with performance comparable to models that use structured EMR data, but we are also able to provide useful rationales to support the predictions, as validated by medical domain expertise. This has important implications for real-world application of explainable clinical decision support from text.

Acknowledgements

Rudzicz is supported by a CIFAR Chair in artificial intelligence.

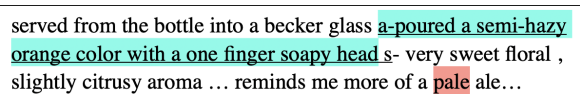
References

- Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. 2018. [Interpretable machine learning in healthcare](#). In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 559–560. ACM.
- Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. [Publicly available clinical BERT embeddings](#). *CoRR*, abs/1904.03323.
- Derek C Angus and Randy S Wax. 2001. [Epidemiology of sepsis: An update](#). *Critical care medicine*, 29(7):S109–S116.
- Michela Assale, Linda Greta Dui, Andrea Cina, Andrea Seveso, and Federico Cabitza. 2019. [The revival of the notes field: Leveraging the unstructured content in electronic health records](#). *Frontiers in Medicine*, 6.
- Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. 2019. [Artificial intelligence, bias and clinical safety](#). *BMJ Qual Saf*, 28(3):231–237.
- Matthew M Churpek, Ashley Snyder, Xuan Han, Sarah Sokol, Natasha Pettit, Michael D Howell, and Dana P Edelson. 2017. [Quick sepsis-related organ failure assessment, systemic inflammatory response syndrome, and early warning scores for detecting clinical deterioration in infected patients outside the intensive care unit](#). *American Journal of Respiratory and Critical Care Medicine*, 195(7):906–911.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. [What does BERT look at? an analysis of BERT’s attention](#). In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.
- Phil Culliton, Michael Levinson, Alice Ehresman, Joshua Wherry, Jay S Steingrub, and Stephen I Galant. 2017. [Predicting severe sepsis using text from the electronic health record](#).
- Thomas Desautels, Jacob Calvert, Jana Hoffman, Melissa Jay, Yaniv Kerem, Lisa Shieh, David Shimabukuro, Uli Chettipally, Mitchell D Feldman, Chris Barton, et al. 2016. [Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach](#). *JMIR Medical Informatics*, 4(3):e28.
- J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). In *NAACL-HLT*.
- Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. 2014. [Unfolding physiological state: Mortality modelling in intensive care units](#). In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84.
- Lauren Gordon, Teodor Grantcharov, and Frank Rudzicz. 2019. [Explainable Artificial Intelligence for Safe Intraoperative Decision Support](#). *JAMA Surgery*, pages 10–11.
- Paulina Grnarova, Florian Schmidt, Stephanie L Hyland, and Carsten Eickhoff. 2016. [Neural document embeddings for intensive care patient mortality prediction](#). *NIPS 2016 Workshop on Machine Learning for Health*.
- Eren Gultepe, Jeffrey P Green, Hien Nguyen, Jason Adams, Timothy Albertson, and Ilias Tagkopoulos. 2013. [From vital signs to clinical outcomes for patients with sepsis: A machine learning basis for a clinical decision support system](#). *Journal of the American Medical Informatics Association*, 21(2):315–325.
- Steven Horng, David A Sontag, Yoni Halpern, Yacine Jernite, Nathan I Shapiro, and Larry A Nathanson. [Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning](#). *PloS ONE*.
- Harold Hotelling. 1992. [Relations between two sets of variates](#). In *Breakthroughs in statistics*, pages 162–190. Springer.
- Sarthak Jain and Byron C. Wallace. 2019. [Attention is not explanation](#). In *NAACL-HLT*.
- 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. 2016. [MIMIC-III, a freely accessible critical care database](#). *Scientific data*, 3:160035.
- Faiza Khan Khattak, Serena Jeblee, Noah Crampton, Muhammad Mamdani, and Frank Rudzicz. 2019. [Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties](#). In *MED-INFO 2019*, pages 1512–1513.

- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, D. Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. [BioBERT: a pre-trained biomedical language representation model for biomedical text mining](#). *Bioinformatics*.
- Tao Lei, R. Barzilay, and T. Jaakkola. 2016. [Rationalizing neural predictions](#). In *EMNLP*.
- Chang Liu, Xinyun Chen, Richard Shin, Mingcheng Chen, and Dawn Song. 2016. Latent attention for if-then program synthesis. In *Advances in Neural Information Processing Systems*, pages 4574–4582.
- Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. [Learning attitudes and attributes from multi-aspect reviews](#). In *2012 IEEE 12th International Conference on Data Mining*, pages 1020–1025. IEEE.
- Ari S. Morcos, Maithra Raghu, and Samy Bengio. 2018. Insights on representational similarity in neural networks with canonical correlation. In *NeurIPS*.
- James Mullenbach, Sarah Wiegrefe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. [Explainable prediction of medical codes from clinical text](#). pages 1101–1111.
- Agnieszka Mykowiecka, Małgorzata Marciniak, and Anna Kupść. 2009. [Rule-based information extraction from patients’ clinical data](#). *Journal of Biomedical Informatics*, 42(5):923–936.
- Shamim Nemati, Andre Holder, Fereshteh Razmi, Matthew D Stanley, Gari D Clifford, and Timothy G Buchman. 2018. [An interpretable machine learning model for accurate prediction of sepsis in the ICU](#). *Critical care medicine*, 46(4):547–553.
- Kimberly J O’Malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. 2005. [Measuring diagnoses: ICD code accuracy](#). *Health services research*, 40(5p2):1620–1639.
- Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pages 6076–6085.
- Sofia Serrano and Noah A. Smith. 2019. [Is attention interpretable?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.
- Yuqi Si and Kirk Roberts. 2019. Deep patient representation of clinical notes via multi-task learning for mortality prediction. *AMIA Summits on Translational Science Proceedings*, 2019:779.
- R Andrew Taylor, Joseph R Pare, Arjun K Venkatesh, Hani Mowafi, Edward R Melnick, William Fleischman, and M Kennedy Hall. 2016. [Prediction of in-hospital mortality in emergency department patients with sepsis: A local big data-driven, machine learning approach](#). *Academic emergency medicine*, 23(3):269–278.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.
- Sarah Wiegrefe and Yuval Pinter. 2019. [Attention is not not explanation](#). pages 11–20.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.
- Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. [Clinical text classification with rule-based features and knowledge-guided convolutional neural networks](#). *BMC Medical Informatics and Decision Making*, 19(3):71.

A Appendix A. On explainability evaluation.

Quantitatively validating latent attention as explanation: As previously noted, evaluating language model explanations is not yet standardized. Despite the effort to make human evaluation fair and reliable, such qualitative measurements are still prone to bias and subjectivity. To validate that latent attention can be used as an explanation, we conduct a stand-alone experiment on the *BeerAdvocate* dataset used by McAuley et al. (2012) and adapted by Lei et al. (2016). This is a dataset that has ground-truth annotations of sentences relevant to prediction results. Although the dataset is not crafted for the purpose of rationale evaluation, we use it as a proxy to examine the quality of our attention scores.



served from the bottle into a becker glass a-poured a semi-hazy orange color with a one finger soapy head s- very sweet floral , slightly citrusy aroma ... reminds me more of a pale ale...

Blue background: attended tokens in annotation
Red background: attended tokens not in annotation
Underscore: annotation

Figure 5: Test case example of BeerAdvocate dataset.

The full *BeerAdvocate* dataset contains 1.5 million beer reviews describing four aspects (i.e., *appearance*, *smell*, *palate*, and *taste*), each corresponding to a rating on a scale of 0 to 5. Lei et al. (2016) published a subset of 90k reviews selected to minimize correlation between *appearance* and other aspects. In our experiment, we use these 90k reviews for training, and 994 annotated reviews for testing. The training set only has rating labels, whereas the testing set has both rating labels and human annotations of sentence-level relevancy. Since all aspects have the exact same setups, it suffices to use the *appearance* rating prediction as a proof-of-concept.

We build a model with only two components, described in Section 3.1, namely BERT (pretrained base-case model) and latent attention. We feed static token embeddings from BERT to a latent attention layer, which output sequence representations to be used for regression through a linear layer with a sigmoid activation. We train the model for 20 epochs and select the best performing one for testing.

In contrast to our clinical model, this model only attends to individual tokens and only generates word-level explanations. For words separated

by the WordPiece tokenizer, we merge the tokens and average the attention weights. For each sentence, we sort the words based on their attention weights and take the top n words as the prediction rationale, where n equals the total length of the human-annotated sentences. We only use attention mechanisms without additional constraints, such as selection continuity, which makes the testing task even more challenging, as the annotations are ranges of words.

The model is evaluated according to mean squared error (MSE) and rationale precision

$$P_{\text{rationale}} = \frac{\sum_{i=1}^N |S_i \cup A_i|}{\sum_{i=1}^N |S_i|},$$

where N is the number of test cases, y is the ground truth rating of appearance, \hat{y} is the predicted rating, A_i is the set of word indices in the annotated covers, S is the set of word indices selected as model explanations, and $|S| = |A|$.

Our model reaches a rationale precision of 76.39%, which indicates that our most attended words are mostly consistent with the annotations. Figure 5 shows an example of *appearance* test results. The experiment demonstrates the usability of latent attention as an explanation mechanism.

Definition of explanation utility in the rating task: For mortality, each sentence is evaluated individually based on how the described situation would contribute to a patient’s survival rate. Sentences describing highly life-threatening complications (such as multiple organ failures) support a positive prediction, whereas sentences indicating improving conditions (such as stable lab measurements) support a negative prediction. In both cases, these sentences are considered helpful. Sentences that are irrelevant (i.e., that support neither a positive nor negative prediction) are considered unhelpful in both populations.

Many of the conditions that present themselves with sepsis onset (such as hypotension) can have numerous etiologies. Diagnostic criteria specify that bacteremia (i.e., bacteria in the bloodstream) must be present in order to predict the development of sepsis. Yet the administration of antibiotics is also not considered as a direct indication of bacteremia without other indications of potential sepsis. Therefore, sentences describing sepsis-related symptoms are not rated as helpful in understanding a positive sepsis prediction until the indication of infection (for example, compromised skin integrity)

Sepsis - Positive Case

... initially admitted on after he was spotted to have partial complex seizure which lasted minutes ... vitals show he was intermittently febrile and hypotensive¹... patient did not feel lightheaded but did appear sleepy ... Pt last had chemo several days prior to admission and is undergoing XRT. Pt currently started on Zosyn, Flagyl for broad coverage pending speciation²... hypotension relieved with IVF. bld cultures check A, urine cultures place central line for access³... review with primary team, pt noted to have bloody bowel movement. GI have already been consulted with the concern being ischaemic colitis given the level of hypotension in the ED⁴... Review of systems is unchanged from admission except as noted below ... doctor aware of bloodstream infection in patient with portacath as he placed portacath bld cultures GPC check⁵, possible pressors less likely now place line if hypotension persists or pressors are initiated ...

Sepsis - Negative Case

... patient lungs clear ... Later BP became hypotensive and HR decreased after receiving neuromuscular reversal agents⁶. Later BP hypertensive with increased pain ... Patient followed commands, afebrile ... Temporary pacer set at backup rate and wires sense and pace... ventilation settings changed to CPAP and resulting ABG wnl, patient was extubated without complications⁷. Temporary atrial sensitivity threshold... No acute distress, regular respiratory, chest expansion symmetric⁸ ... Incision clean dry intact⁹ ... Encourage wakefulness. Increase activity as tolerated...

Mortality - Positive Case

... admit from OR coronary artery bypass graft ... OR course significant for RV failure ... Neuro Arrived sedated ... Resp Poor oxygenation ... Remains on CMV ... Glucose gtt started & titrated. Social updated daughters in on pt condition ... high dose pressors gtt noted ... family decision to withdraw pressors discussed doctors decision made to allow ... bp cont decreasing despite titration ... vent support weaned significantly ... pt became hypotensive with junctional rhythm ... CVP levo and vasopressin weaned significantly today ... UO borderline this am treated with mg iv lasix w/o response. PT started on CVVHD in pm ... dialysate running at cc hr endo pt con at insulin gtt BS s.

Mortality - Negative Case

... Poss extub ... extubated this morning ... wean propofol off. rise to voice, follow commands ... diet advanced to clears. swallows no difficulty. mushroom cath intact w watery melena ... brother visit this eve ... stable for floor transfer ... sat room air sat lungs clear. enc deep breathing and coughing. mobilizing clear secretions. Pt still stable for transfer to floor ... concern re: hypotension but art line shows stable at systolic ... With propofol on board SBP have consistently been ... Lung clear on right, coarse and diminished at left base ... Obese Abdo with active bowel sounds ... Surgery involved. SKIN Intact ... LINES and in place and functioning well. Right radial arterial line also present ...

¹ Description of sepsis-related symptoms which may indicate potential sepsis

² Use of antibiotics indicating possible infections

³ Mentioning blood culture test

⁴ Potential complications of hypotension that may happen in septic patients

⁵ Description of bacteraemia which means high risk of sepsis

⁶ Description non-sepsis-related cause of hypotension

⁷ General good sign of improved patient condition

⁸ No respiratory distress (infection and inflammatory response can cause irregular respiratory rate)

⁹ Presence of incision that is clean and dry, suggesting no infection

Figure 6: Example explanations. Highlighted sentences are rationales picked by our model. Elaboration on the meanings of sentences is written in footnotes. These examples have been edited for increased privacy.

also appears, and vice versa. For negative cases, sentences that are either irrelevant to sepsis or explain other origins of sepsis-related symptoms are rated as helpful. Given this definition, the existence of any helpful sentences means the explanation is valid for a positive case. Similarly, the existence of any unhelpful sentences invalidates a negative case.

Examples of sepsis and mortality explanations are shown in Figure 6. We truncate and edit these texts to avoid data disclosure.