Lessons learned from clinical language processing

Madhumita Sushil

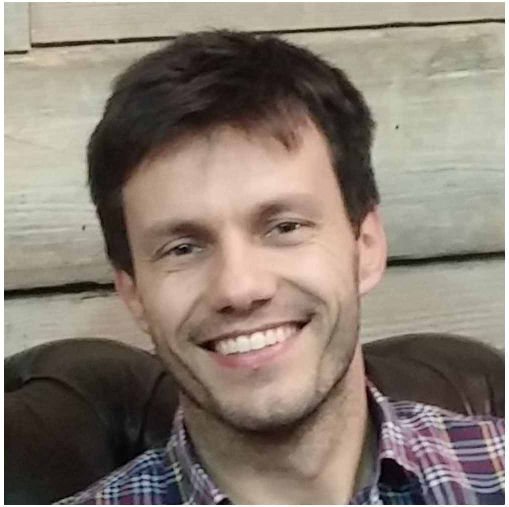
Computational Linguistics & Psycholinguistics Research Center,

University of Antwerp, Belgium

https://madhumitasushil.github.io/

CLiPS
Computational Linguistics & Psycholinguistics
University of Antwerp

Joint work with my advisors



Dr. Simon Šuster
http://simonsuster.github.io/

Prof. Dr. Walter Daelemans
https://www.clips.uantwerpen.be/~walter/

Outline

1. Describing properties of EHR data
   - Non-standard terminology
   - Hedging
   - Imbalance
2. Classifying patient conditions using patient representations
   - Psychiatric symptom severity estimation
   - Task-independent representations
   - Sepsis estimation at end-of-patient-stay
3. Understanding these classifiers
   - Quantifying input feature importance
   - Combining feature importance into patterns
   - Putting important features in context
4. Exploring available domain knowledge in clinical QA, language inference

EHR data characteristics

Non-standard terminology

Hedging

Imbalance

EHR language characteristics: non-standard terminology

A&O
PO
HTN

Abbreviations and acronyms

stress+
stress-

Presence or absence markers

>stress
stress++

Intensity indicators

+/-4 weeks
+-1 week

Approximation

Important to use right tools tailored to medical data.

EHR language characteristics: hedging

EEG showed *no evidence of* seizure.

This finding *might suggest the possibility of* subcortical dysfunction.

Full eye movements horizontally but *seems to have* R gaze preference.

Patient *denies* the use of alcohol.

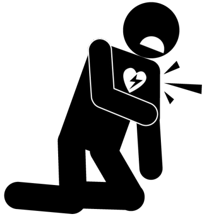
Important to develop tools that can understand nuances of medical language.

EHR data characteristics: imbalance

Imbalance across ethnicities, age groups, diseases, amount of available data.
Important to ensure models aren't biased due to this.

Caucasian,
45 years,
Gastrointestinal disease,
300 records

Caucasian,
60 years,
Cardiovascular disease,
150 records,



Caucasian,
75 years,
Cardiovascular disease,
900 records

Asian,
2 years,
Injury and poisoning,
10 records

Caucasian,
new born,
Perinatal disease,
1 record

Hispanic,
90 years,
Cardiovascular disease,
100 records

Classifying patient conditions

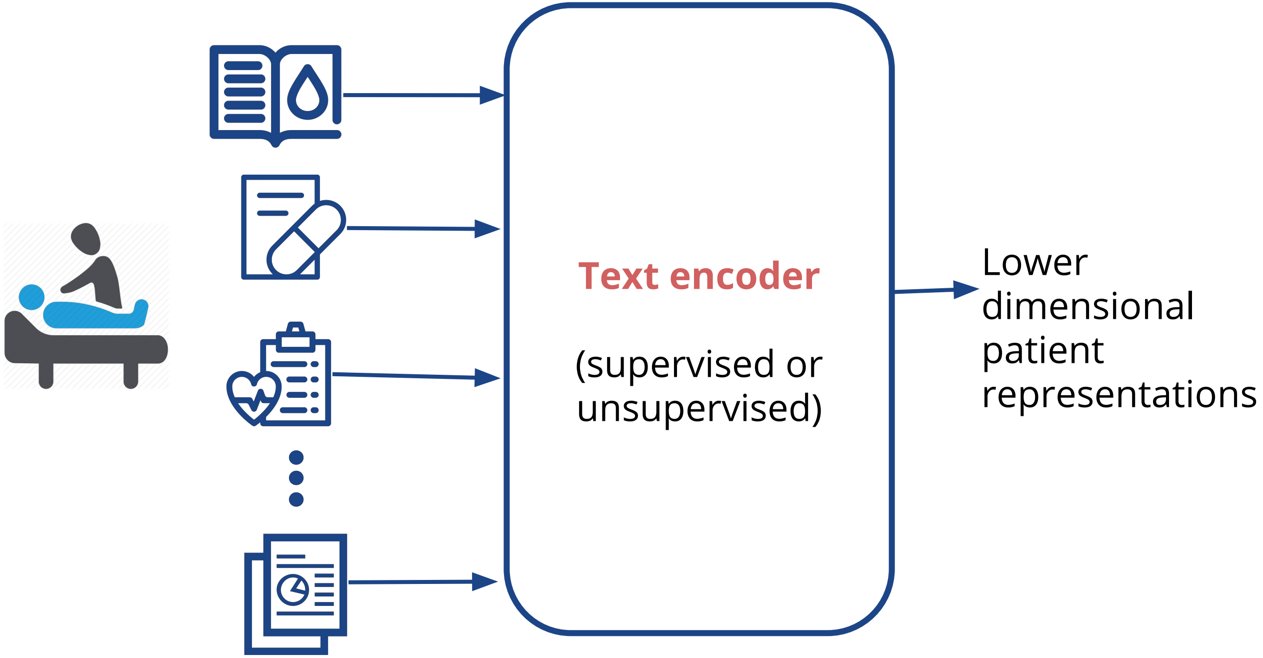
Task dependent patient representations for classifying:

- Psychiatric symptom severity
- Sepsis at end-of-patient-stay

Task independent patient representations for classifying:

- In-hospital, 30 days, 1 year mortality
- Primary diagnostic category
- Primary procedural category

Patient representations



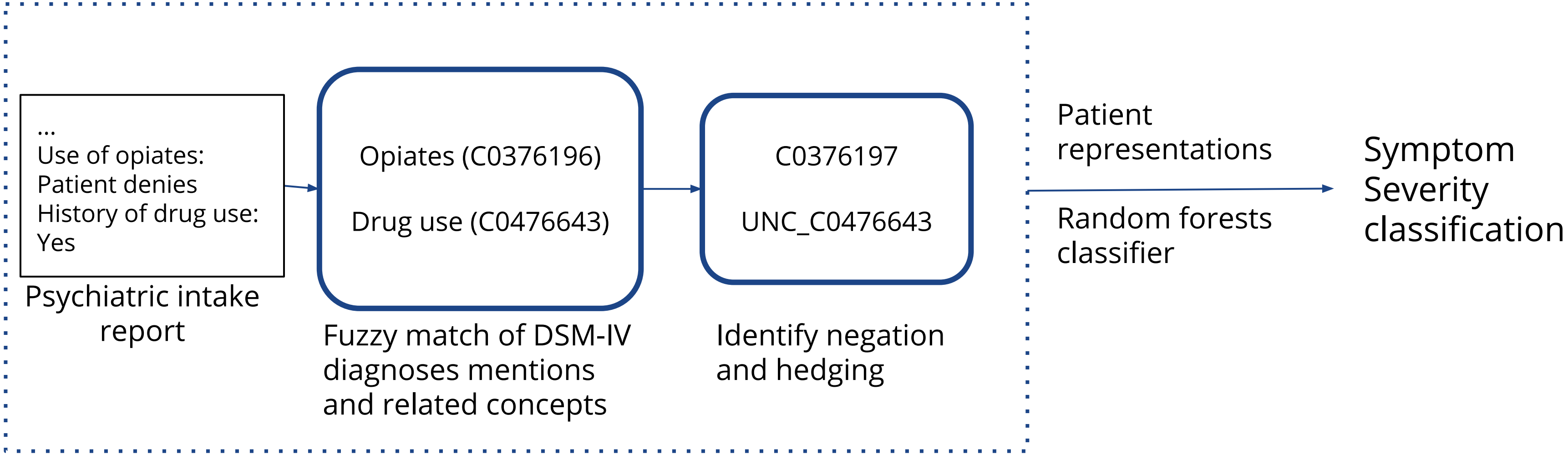
Text encoder

(supervised or unsupervised)

Lower dimensional patient representations

- Encode factual patient info
- Encode text semantics
- Useful for decision-making about patient health

Patient representations for psychiatric symptom severity estimation


…
Use of opiates:
Patient denies
History of drug use:
Yes
Psychiatric intake report
Opiates (C0376196)
Drug use (C0476643)
Fuzzy match of DSM-IV diagnoses mentions and related concepts
C0376197
UNC_C0476643
Identify negation and hedging
Patient representations
Random forests classifier
Symptom Severity classification

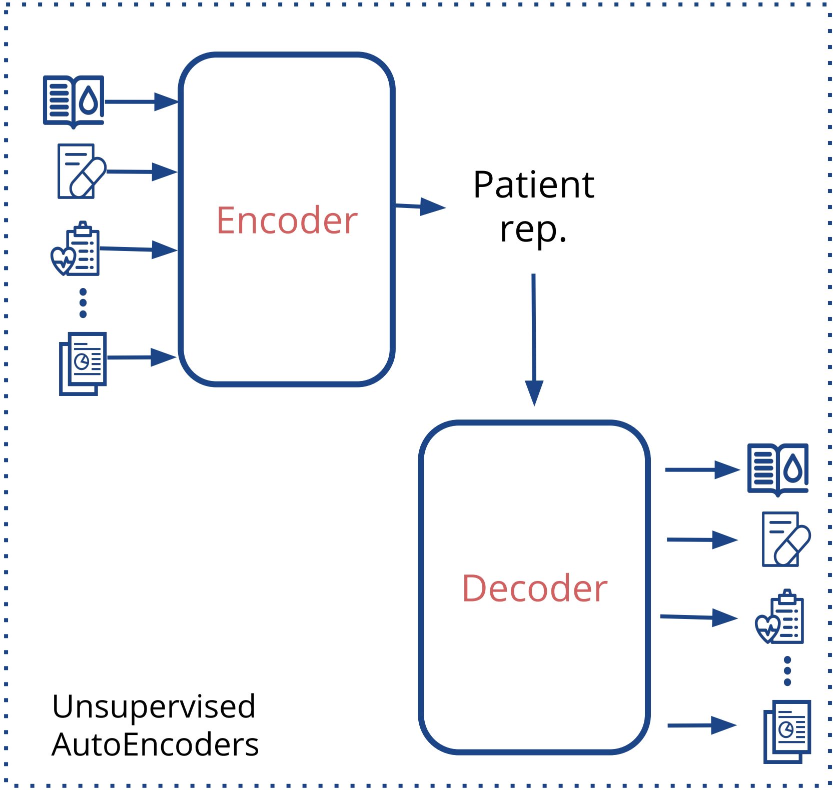
E Scheurwegs, M Sushil, S Tulkens, W Daelemans, and K Luyckx.
Counting trees in random forests: predicting symptom severity in psychiatric intake reports. Journal of Biomedical Informatics, 75: S112-S119, 2017.

Symptom severity classification results

System	10-fold CV (MAE)	Test (MAE)
UMLS concepts (baseline)	72.76 +- 4.42	72.88
UMLS concepts + context	75.49 +- 3.73	79.41
DSM-IV related concepts + context	78.30 +- 2.65	79.52
DSM-IV related concepts + context + self training + outlier removal	**78.77 +- 3.61**	**80.64**

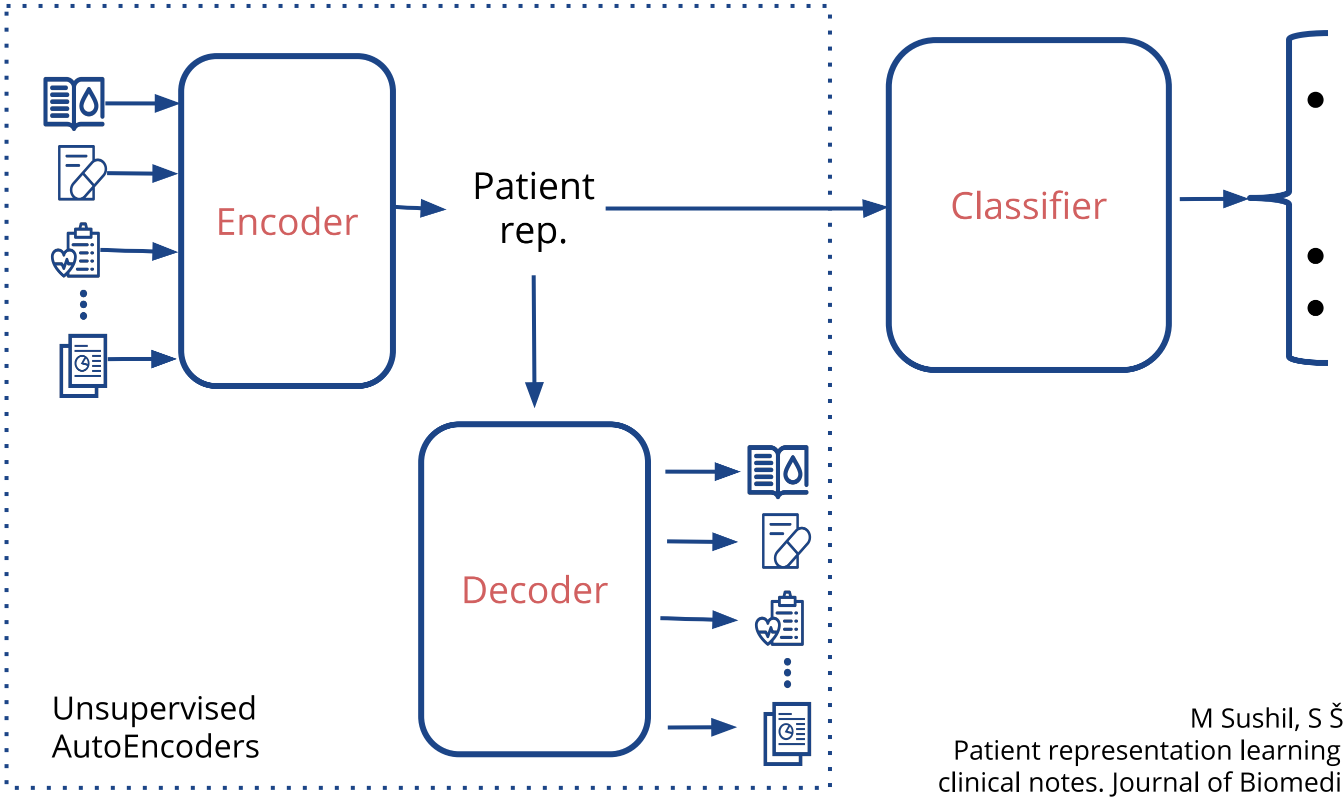
E Scheurwegs, M Sushil, S Tulkens, W Daelemans, and K Luyckx. Counting trees in random forests: predicting symptom severity in psychiatric intake reports. Journal of Biomedical Informatics, 75: S112-S119, 2017.

Task-independent patient representations



M Sushil, S Šuster, K Luyckx, and W Daelemans.
Patient representation learning and interpretable evaluation using clinical notes. Journal of Biomedical Informatics, 84:103 – 113, 2018.

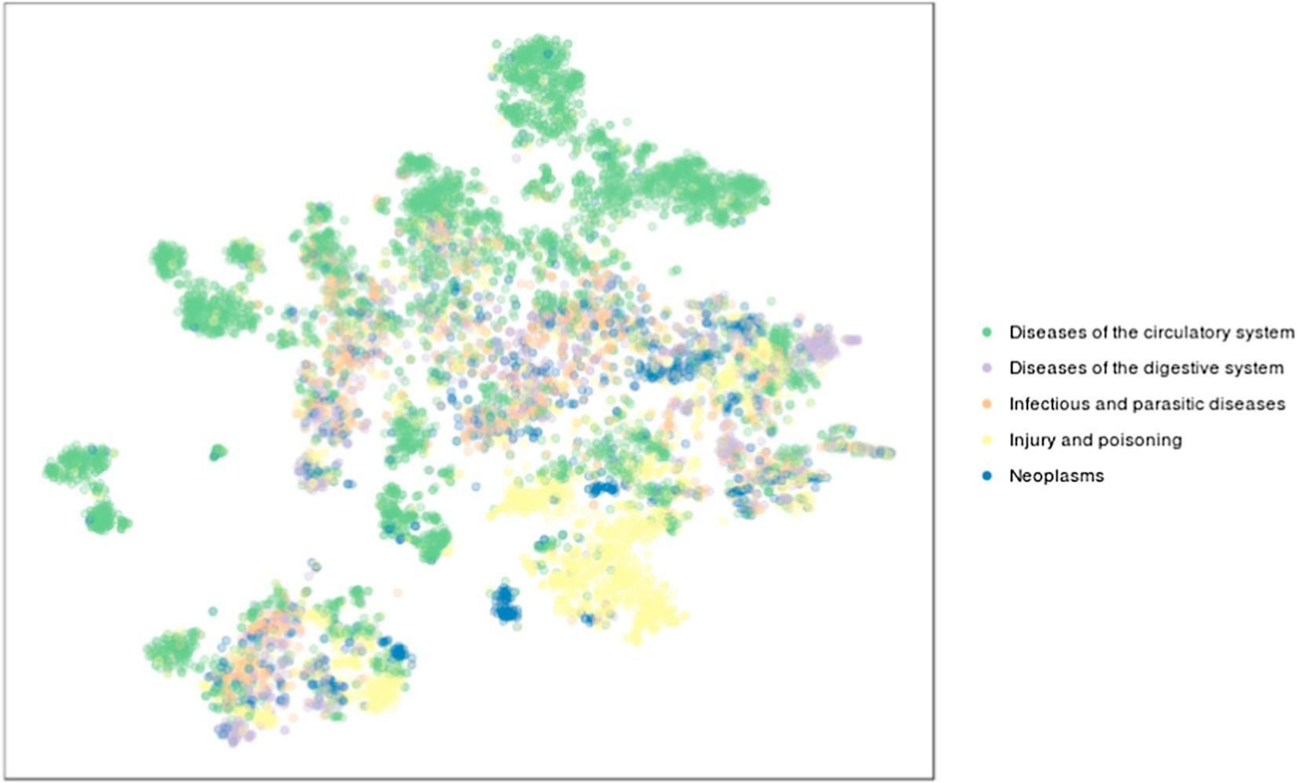
Task-independent patient representations



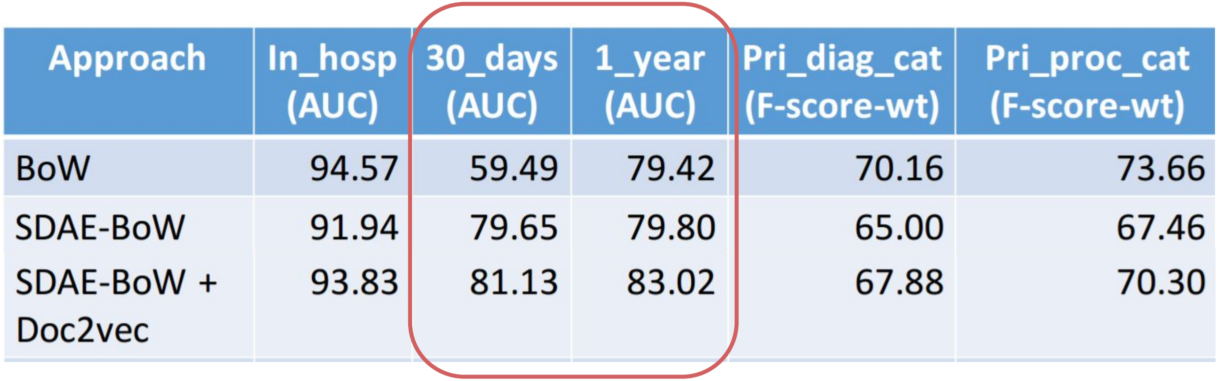
- In hospital/30 days/1 year patient mortality
- Diagnostic category
- Procedural category

M Sushil, S Šuster, K Luyckx, and W Daelemans.
Patient representation learning and interpretable evaluation using clinical notes. Journal of Biomedical Informatics, 84:103 – 113, 2018.

2D visualization of learned representations



Classification performance

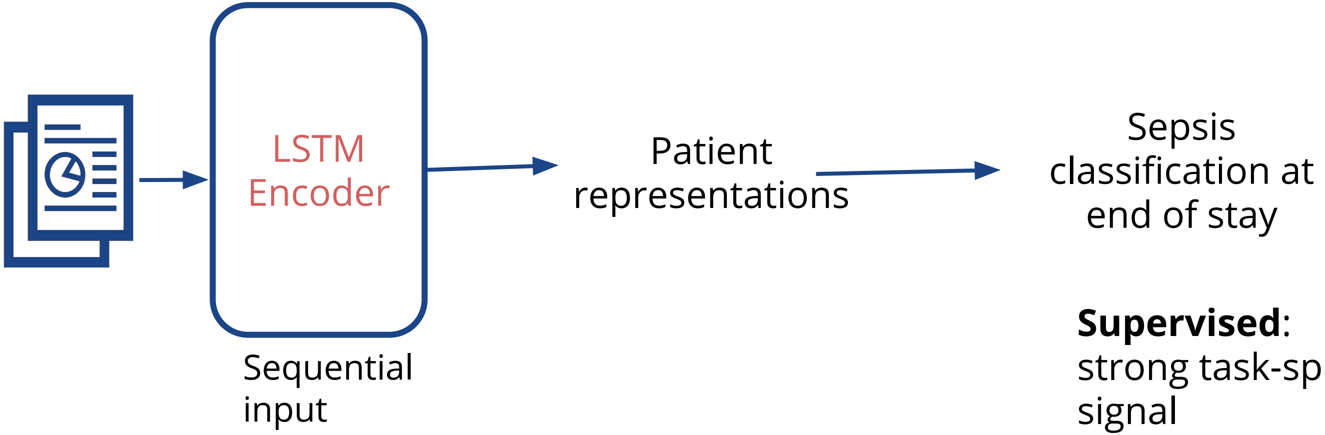


Approach	In_hosp (AUC)	30_days (AUC)	1_year (AUC)	Pri_diag_cat (F-score-wt)	Pri_proc_cat (F-score-wt)
BoW	94.57	59.49	79.42	70.16	73.66
SDAE-BoW	91.94	79.65	79.80	65.00	67.46
SDAE-BoW + Doc2vec	93.83	81.13	83.02	67.88	70.30

Generalized patient representations outperform sparse models for post-discharge mortality prediction where no. of death instances is low.

M Sushil, S Šuster, K Luyckx, and W Daelemans.
Patient representation learning and interpretable evaluation using clinical notes. Journal of Biomedical Informatics, 84:103 – 113, 2018.

Sequential representations for sepsis prediction



LSTM Encoder

Sequential input

Patient representations

Sepsis classification at end of stay

Supervised:
strong task-specific signal

Classification performance

Input	Macro F1	Sepsis F1
Discharge note	0.68	0.41
Last note before discharge	0.60	0.27

Sepsis estimation from clinical notes is a difficult task!

Conclusions

Task-independent patient representations generalize and retain important information across several tasks, which can be used to find soft patient cohorts.

They are promising for tasks with low prevalence due to high data imbalance.

When high performance on one specific end task is the goal, task-specific models can be better.

What have these models learned?

Quantifying feature importance in NN

Moving to important patterns

Using embeddings as input

Accounting for word order in explanations

Need to understand and explore our models

- How can we improve our models?
- Is the model generalized for use across populations?
  - Is it biased towards a specific cohort of patients in one hospital?
  - Is it biased towards properties of the EHR the hospital used?
  - Is it biased towards data pre-processing steps?

Quantifying feature importance in neural networks: Sensitivity Analysis

Quantifying how does changing the input affect the output across 2 networks

$$Saliency_{w_i}^{(j)} = \underbrace{\frac{\partial o_k^{(j)}}{\partial R_i^{(j)}}}_{\text{Sensitivity of output to patient representations}} * \underbrace{\frac{\partial R_i^{(j)}}{\partial w_i^{(j)}}}_{\text{Sensitivity of patient representations to input}}$$

Take mean square value across all instances.

Most important features: sensitivity analysis

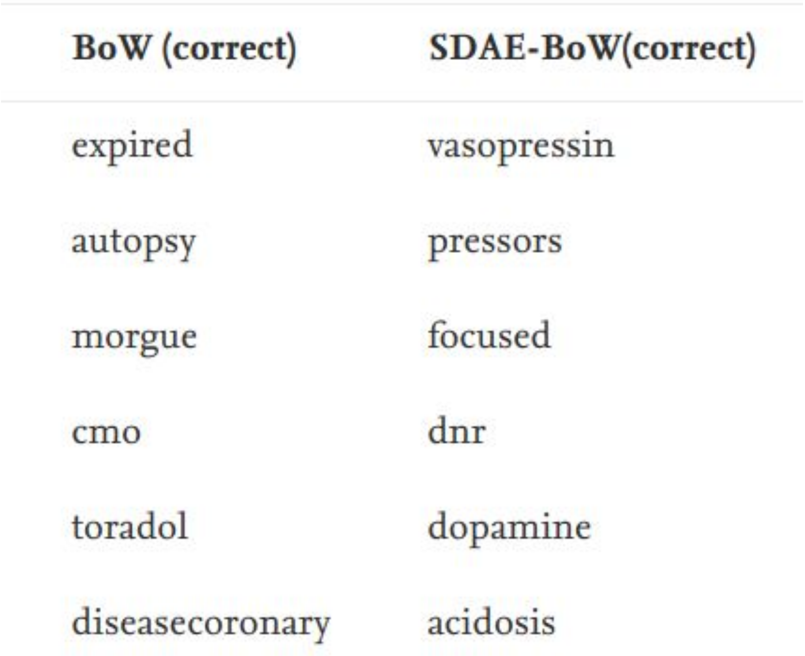
In hospital death	30 days post-discharge death	1 year post-discharge death	Diagnostic category	Procedural category
vasopressin	leaflet	magnevist	NUM	NUM
pressors	structurally	signal	previous	no
focused	sda	decisions	rhythm	of
dnr	periventricular	periventricular	no	enzymes
dopamine	excursion	embolus	flexure	extubated

Several important terms related to patient conditions and treatments.

Absence of terms (in red) often used to rule out certain outputs.

Requires more context to disambiguate use of negation markers, NUM, function words.

Comparing important feature sets



BoW (correct)	SDAE-BoW(correct)
expired	vasopressin
autopsy	pressors
morgue	focused
cmo	dnr
toradol	dopamine
diseasecoronary	acidosis

Bag-of-words sparse supervised representations uses several task-specific important keywords.

Autoencoder-based dense representations focus more on holistic patient view.

Moving to important patterns

If-then-else rule lists:

if ***<condition1>*** and ***<condition2>*** and … ⇒ **class1**

elif ***<condition3>*** … ⇒ **class1**

else **class2**

Quantifies associations between features and classes

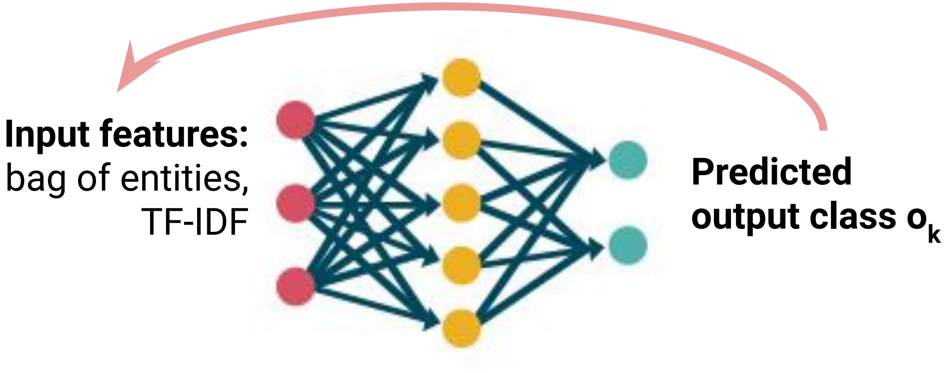
If-then-else rules to interpret neural nets

Input features:
bag of entities,
TF-IDF

Predicted output class o_k

If-then-else rules to interpret neural nets

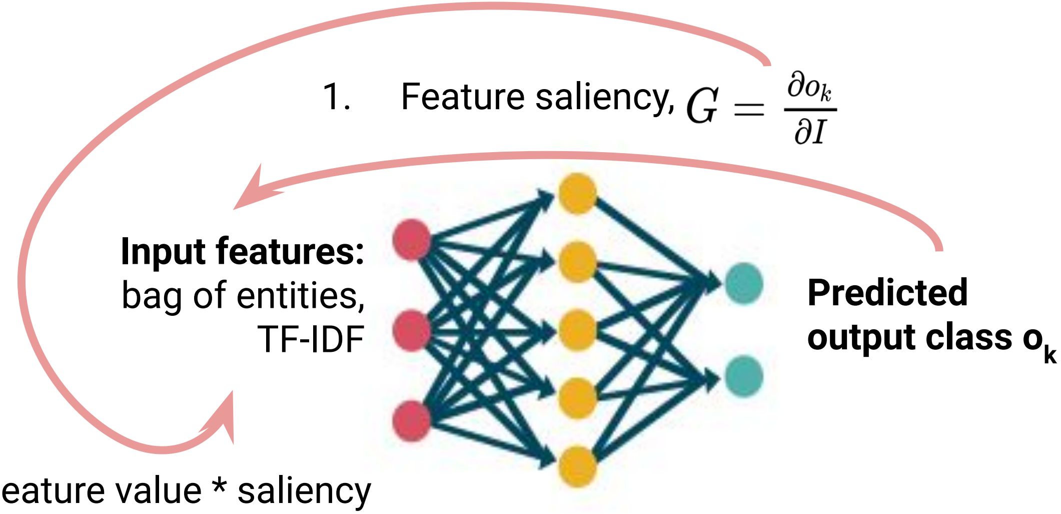
1. Feature saliency, $G = \frac{\partial o_k}{\partial I}$



Input features:
bag of entities,
TF-IDF

Predicted
output class o_k

If-then-else rules to interpret neural nets



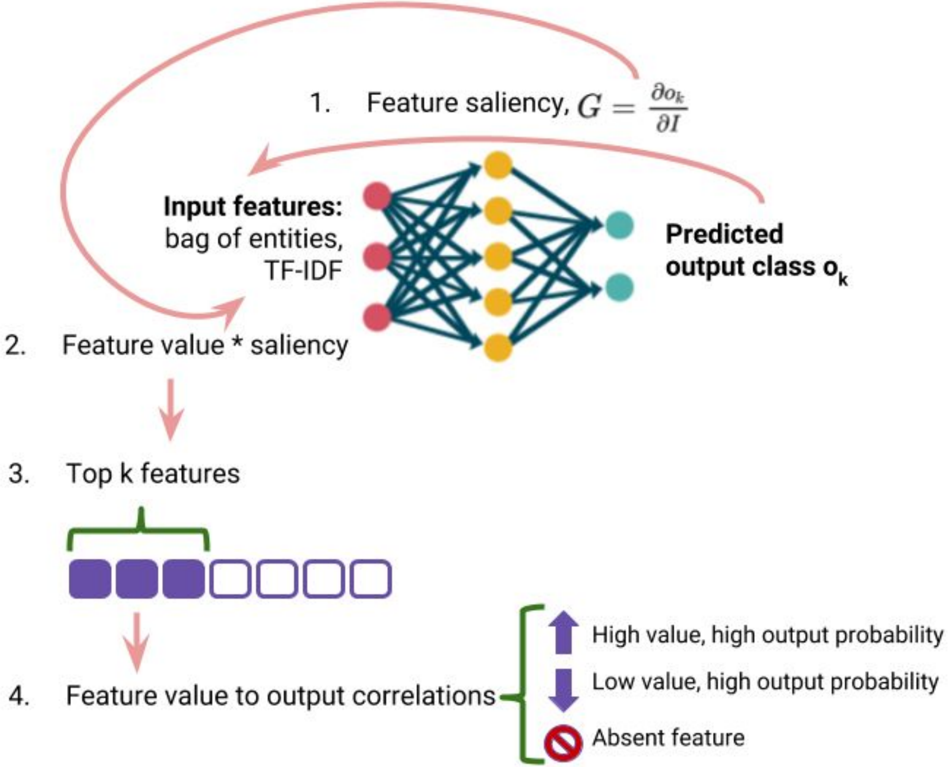
1. Feature saliency, $G = \frac{\partial o_k}{\partial I}$

Input features: bag of entities, TF-IDF

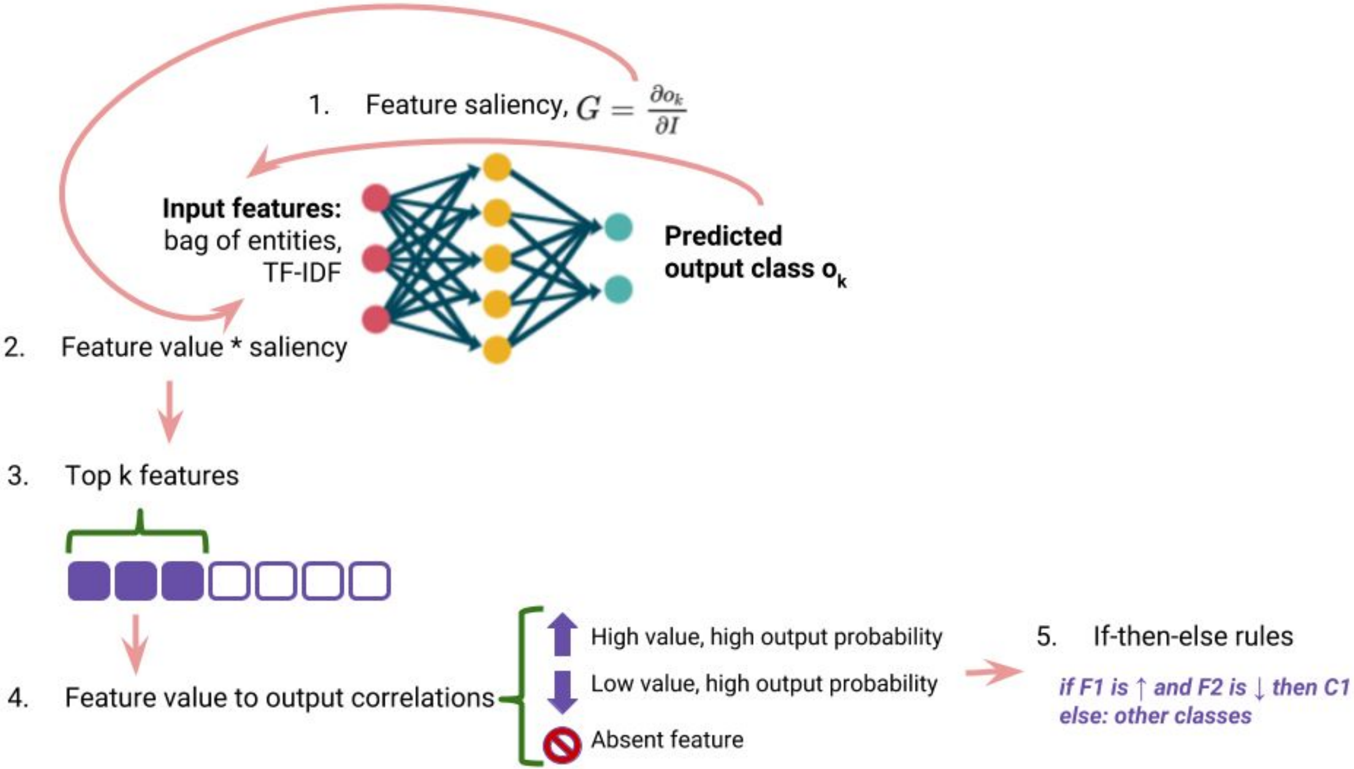
Predicted output class o_k

2. Feature value * saliency

If-then-else rules to interpret neural nets


1. Feature saliency, $G = \frac{\partial o_k}{\partial I}$
Input features:
bag of entities,
TF-IDF
Predicted
output class o_k
2. Feature value * saliency
3. Top k features
4. Feature value to output correlations
High value, high output probability
Low value, high output probability
Absent feature

If-then-else rules to interpret neural nets


1. Feature saliency, $G = \frac{\partial o_k}{\partial I}$
Input features: bag of entities, TF-IDF
Predicted output class o_k
2. Feature value * saliency
3. Top k features
4. Feature value to output correlations
High value, high output probability
Low value, high output probability
Absent feature
5. If-then-else rules
if F1 is ↑ and F2 is ↓ then C1
else: other classes

M Sushil, S Šuster, and W Daelemans. Rule induction for global explanation of trained models. Workshop on Analyzing and interpreting neural networks for NLP (BlackboxNLP), EMNLP 2018

Explanations for primary diagnostic category prediction

↑ **Take blood pressure (treatment)** and
🚫 **Nothing by mouth** and
🚫 **Flagyl**

→**Diseases of the circulatory system** (✅84/90)

Explanations for primary diagnostic category prediction

↑ **Pneumonia** and

↑ **Lung opacity** and

↓ **Non-specific ST-T changes by ECG** and

🚫 **CT of pelvis w/o contrast**

→**Diseases of the respiratory system** (✓7/7)

Explanation for in-hospital mortality prediction

↑ **Physical examination** and

↑ **Pregnancy with medical condition**

→**Dies within hospital** (✅221/222)

Using uninterpretable embeddings as input

Word embeddings as input encode multiple dimensions for every word.

To obtain overall word saliency, instead of saliency over individual dimensions, pool embedding importance scores.

$$saliency_{w_i} = \Sigma_{dim}(emb_{w_i} \odot grad_{dim})$$

M Sushil, S Šuster, and W Daelemans. Distilling neural networks into skipgram level decision lists. Computing Research Repository, 2005.07111, 2020.

Accounting for word order in explanations

Find most-important skipgrams instead of individual words.

*"**no** signs **of infection** were found. "*

$$saliency_{no\ of\ infection} = \frac{saliency_{no} + saliency_{of} + saliency_{infection}}{3}$$

M Sushil, S Šuster, and W Daelemans. Distilling neural networks into skipgram level decision lists. Computing Research Repository, 2005.07111, 2020.

Explanations for sepsis prediction classifier

↑ **sepsis major surgical** →**septic** (✔209/209)

🚫 **complaint : sepsis** and

↑ **chief hypotension major** →**septic** (✔169/169)

Explanations for sepsis prediction classifier

↑ **indication endocarditis .** →**septic** (✔34/34)

🚫 **day ventilation** and
🚫 **rhythm . low lead** and
🚫 **sepsis ; _** and
🚫 **pmicu nursing progress** and
🚫 **indication endocarditis .** and
↑**admitting sepsis** and
🚫 **reason : of** and
🚫 **3 ,** →**septic** (✔103/113)

Conclusions

Finding patterns learned by models provides insights into its internal working.

While in some cases learned patterns correspond medical knowledge, in other cases models also pick up on the biases in the dataset due to small size or task formulation.

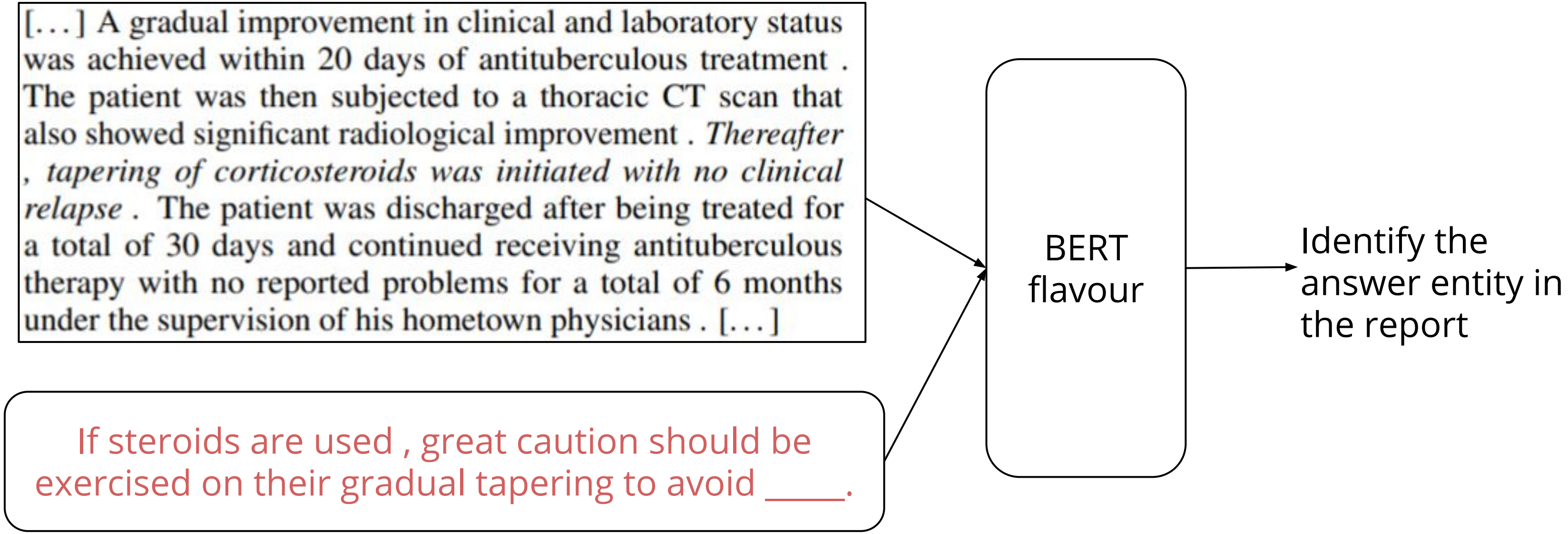
Our illustrations reinforce the benefits of understanding trained models for both improving them and removing biases in them.

Exploring use of domain knowledge in clinical NLP

Clinical case reports for question answering

Medical language inference

Clinical case reports for question answering



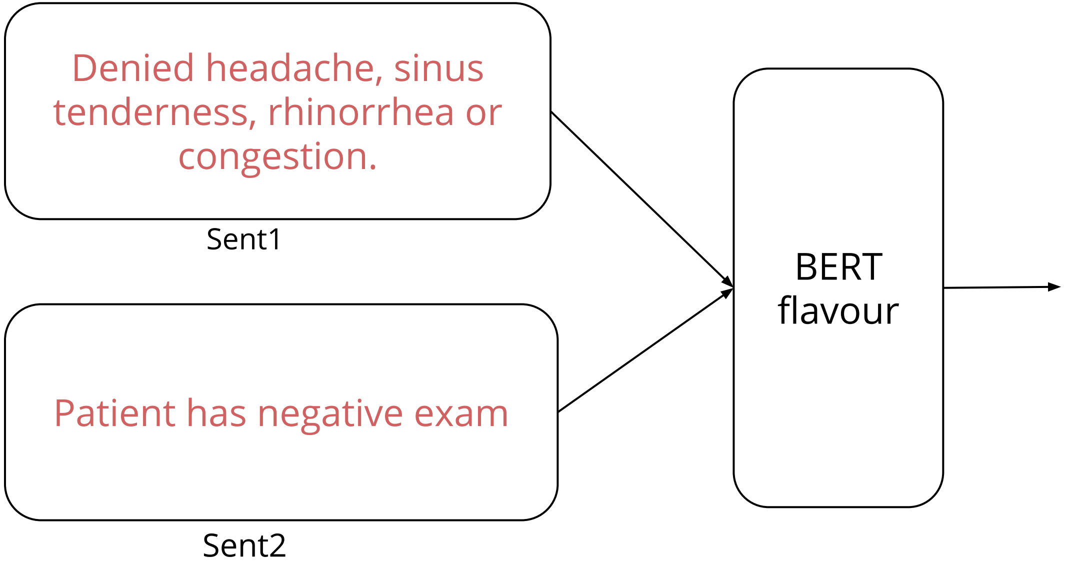
[. . .] A gradual improvement in clinical and laboratory status was achieved within 20 days of antituberculous treatment . The patient was then subjected to a thoracic CT scan that also showed significant radiological improvement . *Thereafter , tapering of corticosteroids was initiated with no clinical relapse* . The patient was discharged after being treated for a total of 30 days and continued receiving antituberculous therapy with no reported problems for a total of 6 months under the supervision of his hometown physicians . [. . .]

If steroids are used , great caution should be exercised on their gradual tapering to avoid _____.

Question answering results

	F1	Requiring domain knowledge (74)
Human (expert) (subset)	**53.7**	**60**
BERT-base-cased	43.6	45
BERT-base-cased +Pubmed 1M	48.3	48
BERT-base-cased +Pubmed 1M + MedNLI	48.7	51

Medical Language Inference



Denied headache, sinus tenderness, rhinorrhea or congestion.

Sent1

Patient has negative exam

Sent2

BERT flavour

Is the second sentence implied by the first one?

Yes (Entailment)/
No (Contradiction)/
Neutral

Medical language inference

	F1
InferSent	76.0
BERT-base-cased	81.0
BERT-base-cased + Pubmed 1M	83.9

Error analysis - Medical language Inference

Error type	Definite	Probable
Incorrect negation	1	2
Incorrect temporal resolution		2
No domain knowledge	11	4
Abbreviation resolution		1
Lack of common sense	1	
Assumption of missing info	1	
Difficult cases	2	1
Incorrect/conflicting annotation	1	

Medical NLI error analysis - BERT+PubMed

No history of blood clots or DVTs, has never had **chest pain** prior to one week ago.

Patient has **angina**

Entailment -> contradiction

Her a[** Location **]e and **PO** intake have been normal.

She has been **NPO** since midnigh

Contradiction -> neutral

Medical NLI error analysis - BERT+PubMed

HISTORY OF PRESENT ILLNESS: A 34-year-old male status post high speed motor vehicle **crash** unrestrained driver.

Patient has recent **trauma**

Entailment -> neutral

Infusion stopped and she was treated with **Benadryl** 50 mg x 1, prednisone 40 mg x 1, ativan 1 mg.

Patient has had an **allergic reaction**

Entailment -> neutral

Next steps: Incorporating domain knowledge

Lack of domain knowledge limits capabilities of text understanding in existing systems.

Explicitly incorporating domain knowledge from medical textbooks, encyclopedias would improve NLU.

Directions for future

Combining multiple modalities

Clinical notes are rich, but provide incomplete information.

Jointly utilizing clinical notes, time-series and other structured data, and imaging data can improve patient outcome estimation.

Integrating explanations within model structure

Post-hoc explanations, while useful, are not 100% accurate.

Developing models which output explanations jointly during classification would increase transparency.

Developing models generalizable across populations

Several sources of biases are present in EHR data.

Increasingly important to identify existing biases in a model and remove them.

Causal inference of patient outcomes

Current systems frequently exploit correlations.

Moving towards causal reasoning, as opposed to correlation-based reasoning:

can improve capabilities and find new clinical hypotheses for testing.

might make models inherently interpretable.

References

- E Scheurwegs, M Sushil, S Tulkens, W Daelemans, and K Luyckx. Counting trees in random forests: predicting symptom severity in psychiatric intake reports. Journal of Biomedical Informatics, 75: S112-S119, 2017.
- M Sushil, S Šuster, K Luyckx, and W Daelemans. Patient representation learning and interpretable evaluation using clinical notes. Journal of Biomedical Informatics, 84:103 – 113, 2018.
- M Sushil, S Šuster, and W Daelemans. Rule induction for global explanation of trained models. Workshop on Analyzing and interpreting neural networks for NLP (BlackboxNLP), EMNLP 2018
- M Sushil, S Šuster, and W Daelemans. Distilling neural networks into skipgram level decision lists. Computing Research Repository, 2005.07111, 2020.
- Icons
  - Health Care by Komkrit Noenpoempisut from the Noun Project
  - https://www.freeiconspng.com/img/9249
  - https://www.kindpng.com/imgv/iwoRbim_cardiovascular-disease-myocardial-infarction-heart-heart-attack-icon/
  - https://icon-library.net/icon/old-icon-28.html
  - Patient by Delwar Hossain from the Noun Project
  - Blood donation registry by pictohaven from the Noun Project
  - Prescription by LAFS from the Noun Project
  - Ecg Report by ProSymbols from the Noun Project
  - Question by Elves Sousa from the Noun Project