# Synthetic dataset to explain and evaluate rules learned by RNNs

**Madhumita Sushil,** Simon Šuster, Walter Daelemans

CLiPS
Computational Linguistics & Psycholinguistics
University of Antwerp

# Interpretable Machine Learning

WHAT

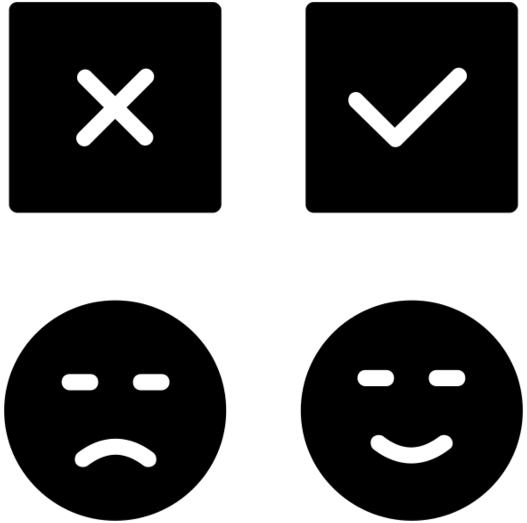
TO WHOM

Providing **explanations** **to humans** to facilitate them **to understand the cause of a model's decision**

WHY

Miller, Tim. “Explanation in artificial intelligence: Insights from the social sciences.” arXiv Preprint arXiv:1706.07269. (2017).

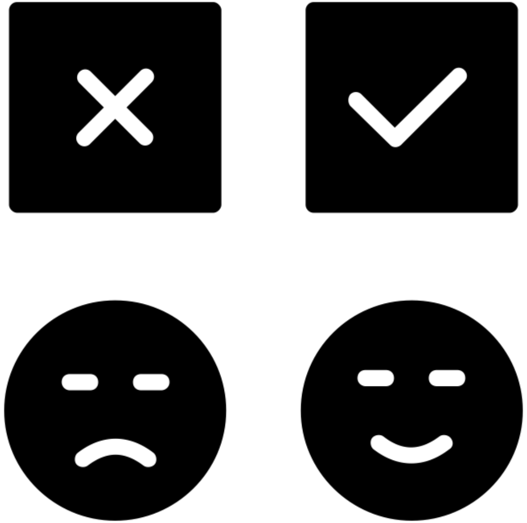
# Is my explanation valid?



**Reliance on human evaluation**

close ended survey by ProSymbols from the Noun Project

# Is my explanation valid?

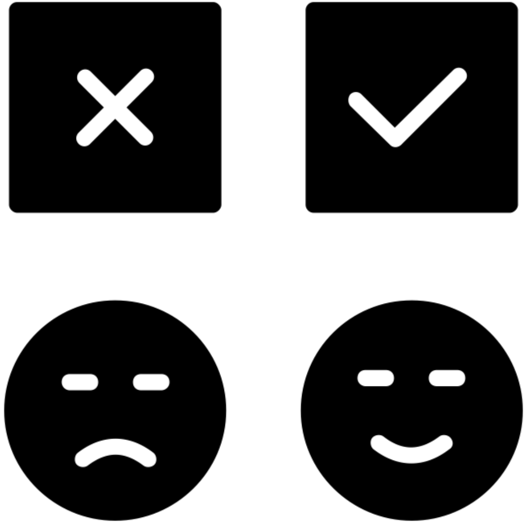


Reliance on human evaluation

**It may not be the only way to solve a task**

close ended survey by ProSymbols from the Noun Project

# Is my explanation valid?



Reliance on human evaluation

**Not always available in complex domains**

close ended survey by ProSymbols from the Noun Project

# Synthetic data for controlled evaluation

- Should model real corpora.

- Should have a predetermined labeling pattern for explanation evaluation.

# Domain guided sentence sampling (MIMIC-III)

***Infection*** **keywords**

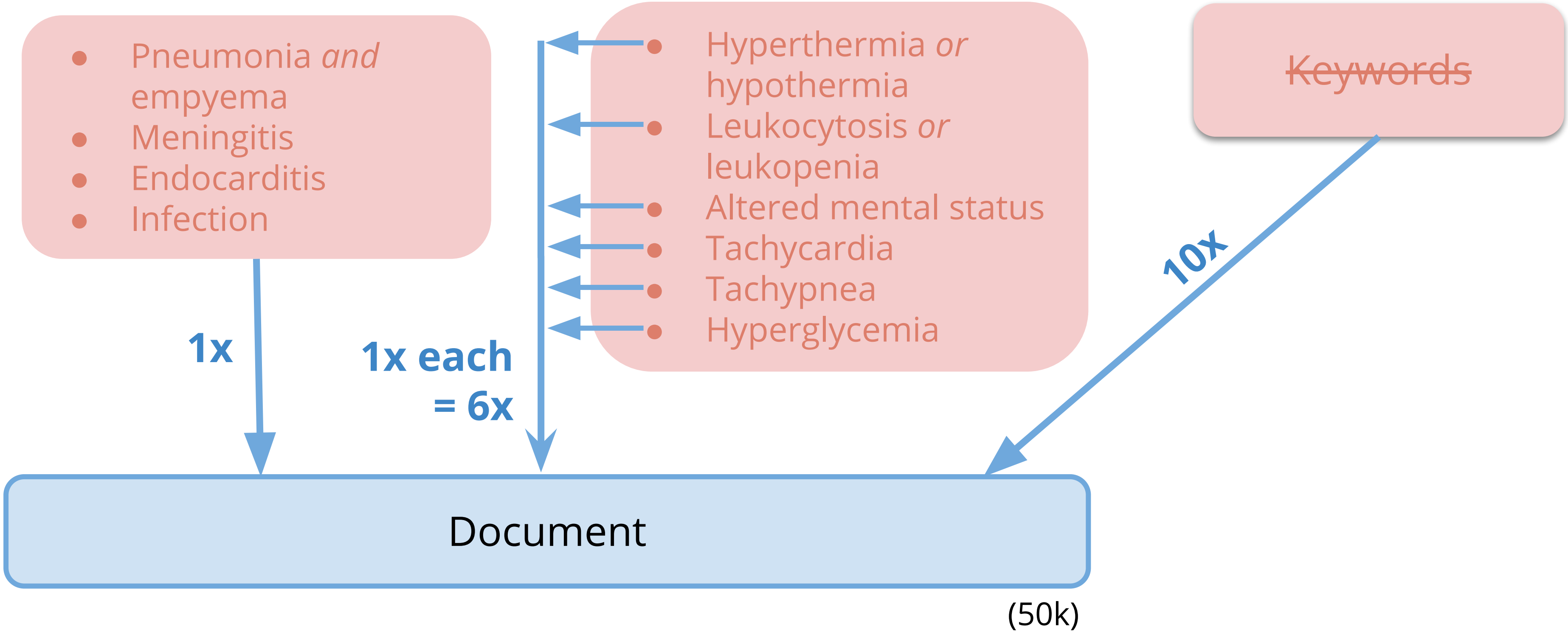
- Pneumonia *and* empyema
- Meningitis
- Endocarditis
- Infection

***Measure*** **keywords**

- Hyperthermia *or* hypothermia
- Leukocytosis *or* leukopenia
- Altered mental status
- Tachycardia
- Tachypnea
- Hyperglycemia

**~~Keywords~~**

# Populating documents



- Pneumonia *and* empyema
- Meningitis
- Endocarditis
- Infection

1x

- Hyperthermia *or* hypothermia
- Leukocytosis *or* leukopenia
- Altered mental status
- Tachycardia
- Tachypnea
- Hyperglycemia

1x each = 6x

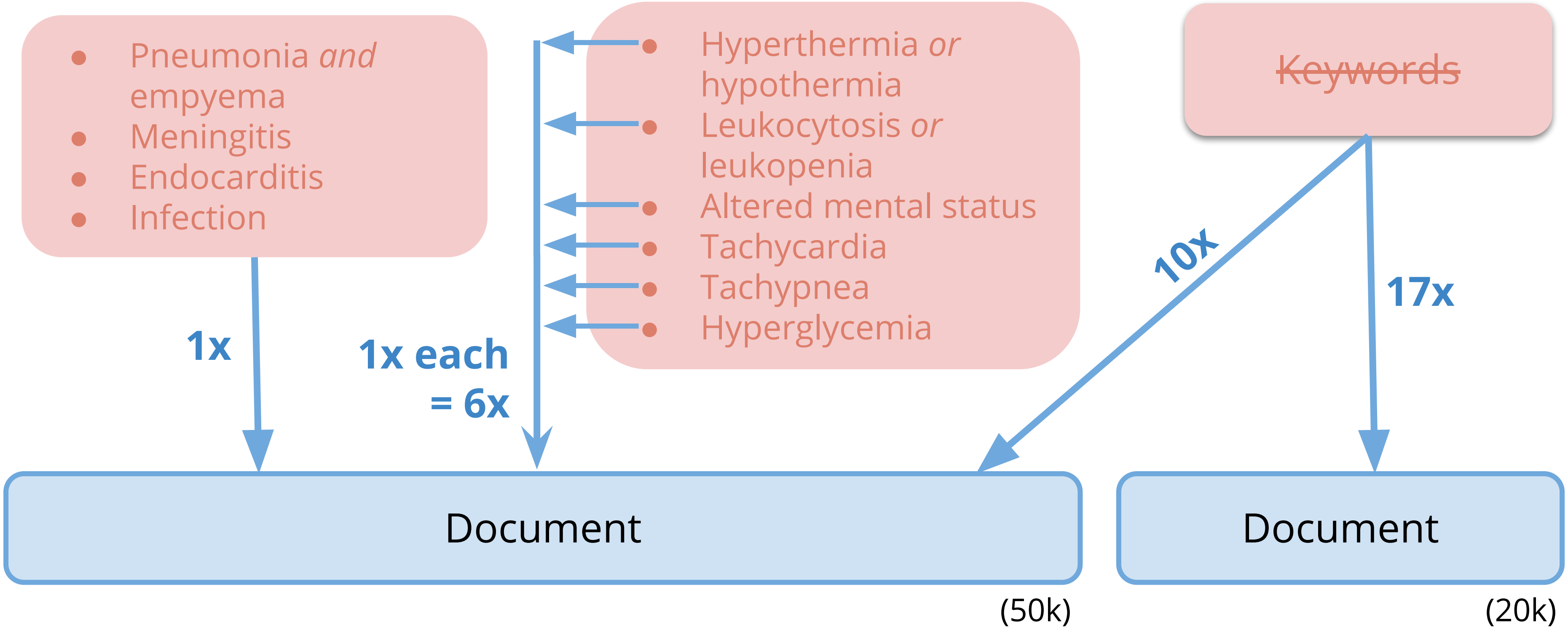
~~Keywords~~

10x

Document

(50k)

# Populating documents



- Pneumonia *and* empyema
- Meningitis
- Endocarditis
- Infection

- Hyperthermia *or* hypothermia
- Leukocytosis *or* leukopenia
- Altered mental status
- Tachycardia
- Tachypnea
- Hyperglycemia

~~Keywords~~

# Labeling documents

1. Find if keyword terms are negated

2. Label according to rule:

If ***infection_keyword*** is ***not negated***

and ***at least 2 measure_keywords*** are ***not negated***:

class label: *'septic'* (class A)

*'non-septic'* otherwise (class notA)

# Dataset statistics

**Class distribution:** 49% sepsis (class A)

**Vocabulary size:** 47015

# Gold important terms - document level

## Mentions of

- pneumonia
- empyema
- meningitis
- endocarditis
- infection
- hyperthermia
- hypothermia
- leukocytosis
- leukopenia
- altered
- mental
- status
- tachycardia
- tachypnea
- hyperglycemia

+ **Corresponding negation markers**

**No evidence of** **infection** was found.
**Altered mental status** exists.

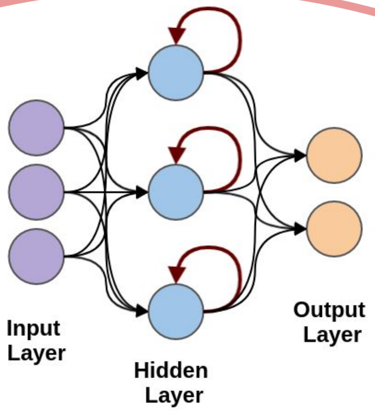
# Classifiers to explain - LSTM (Macro-F)

| Model/ Embedding | LSTM 100d | LSTM 50d |
|---|---|---|
| Embedding 100d | **0.97** | 0.92 |
| Embedding 50d | 0.96 | 0.92 |

# Explaining RNNs - Pipeline

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

Word sequence embeddings



Predicted output class $o_k$

2. Computing word importance

3. Identifying top skipgrams

🟪 □ 🟪 🟪 🟧 🟧 instance1

🟥 □ 🟧 🟧 🟩 🟩 🟩 instance2

4. Discretize skipgram importance

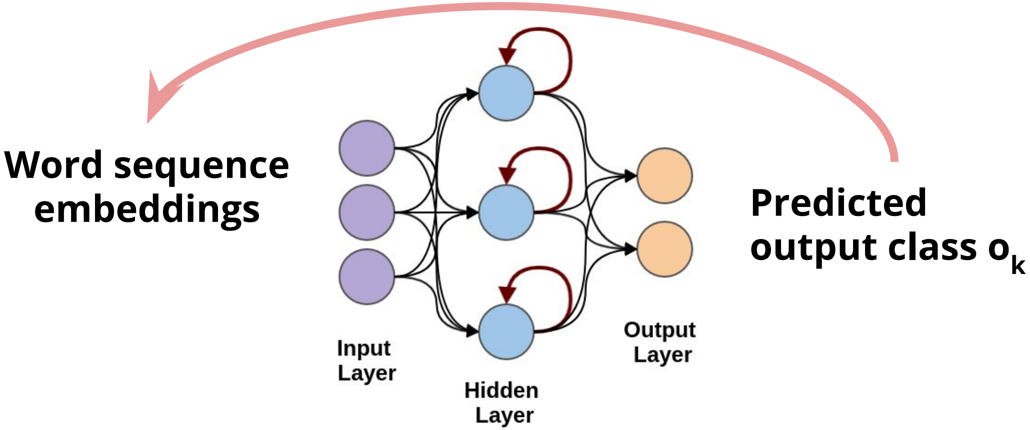
- + Positive impact on output probability
- − Negative impact on output probability
- 🚫 Absent in input sequence

5. Rules as explanations

***if F1 is + and F2 is - then C1***
***else: C2***

# 1. Input saliency

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



**Word sequence embeddings**

**Predicted output class o$_k$**

RNN image borrowed from [https://austingwalters.com/classify-sentences-via-a-recurrent-neural-network-lstm/](https://austingwalters.com/classify-sentences-via-a-recurrent-neural-network-lstm/)

# 2. Gradient pooling for word importance

## Using gradients only

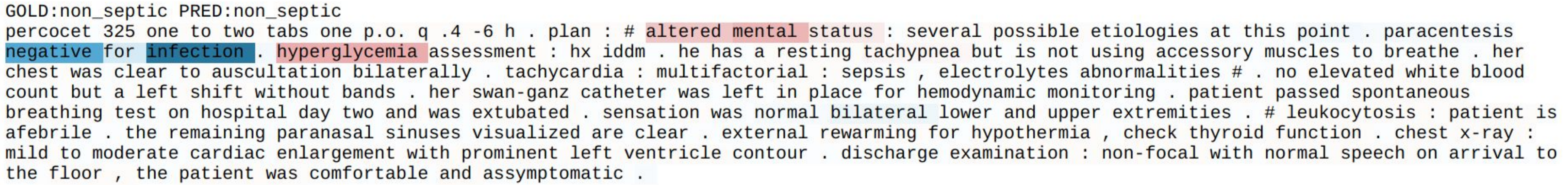
- Sum: $\mathbf{\Sigma}_{\text{dim}}\ grad$
- L2 norm: $\mathbf{\Sigma}_{\text{dim}}\ grad^2$

## Using gradients + embeddings

- Dot product: $\mathbf{\Sigma}_{\text{dim}}\ (emb \odot grad)$

# Qualitative comparison: Heatmaps

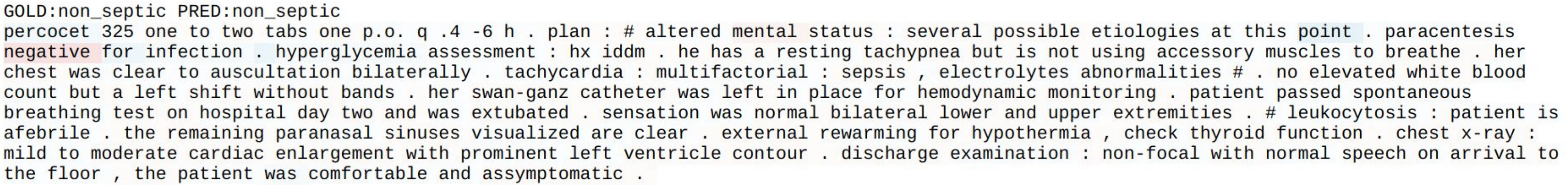
**dot**:



GOLD:non_septic PRED:non_septic

percocet 325 one to two tabs one p.o. q .4 -6 h . plan : # altered mental status : several possible etiologies at this point . paracentesis negative for infection . hyperglycemia assessment : hx iddm . he has a resting tachypnea but is not using accessory muscles to breathe . her chest was clear to auscultation bilaterally . tachycardia : multifactorial : sepsis , electrolytes abnormalities # . no elevated white blood count but a left shift without bands . her swan-ganz catheter was left in place for hemodynamic monitoring . patient passed spontaneous breathing test on hospital day two and was extubated . sensation was normal bilateral lower and upper extremities . # leukocytosis : patient is afebrile . the remaining paranasal sinuses visualized are clear . external rewarming for hypothermia , check thyroid function . chest x-ray : mild to moderate cardiac enlargement with prominent left ventricle contour . discharge examination : non-focal with normal speech on arrival to the floor , the patient was comfortable and assymptomatic .

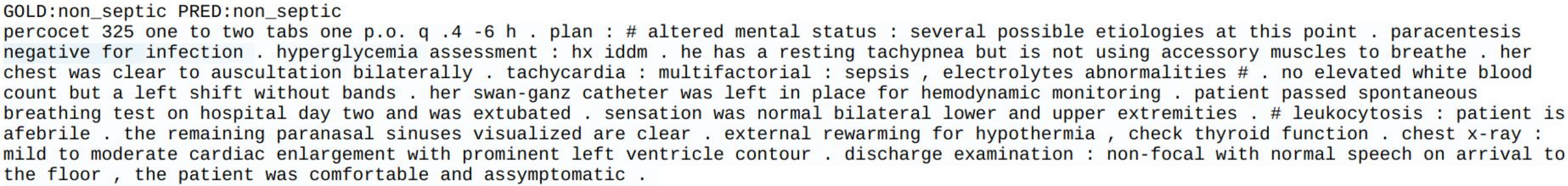
sum:



GOLD:non_septic PRED:non_septic

percocet 325 one to two tabs one p.o. q .4 -6 h . plan : # altered mental status : several possible etiologies at this point . paracentesis negative for infection . hyperglycemia assessment : hx iddm . he has a resting tachypnea but is not using accessory muscles to breathe . her chest was clear to auscultation bilaterally . tachycardia : multifactorial : sepsis , electrolytes abnormalities # . no elevated white blood count but a left shift without bands . her swan-ganz catheter was left in place for hemodynamic monitoring . patient passed spontaneous breathing test on hospital day two and was extubated . sensation was normal bilateral lower and upper extremities . # leukocytosis : patient is afebrile . the remaining paranasal sinuses visualized are clear . external rewarming for hypothermia , check thyroid function . chest x-ray : mild to moderate cardiac enlargement with prominent left ventricle contour . discharge examination : non-focal with normal speech on arrival to the floor , the patient was comfortable and assymptomatic .

L2:



GOLD:non_septic PRED:non_septic

percocet 325 one to two tabs one p.o. q .4 -6 h . plan : # altered mental status : several possible etiologies at this point . paracentesis negative for infection . hyperglycemia assessment : hx iddm . he has a resting tachypnea but is not using accessory muscles to breathe . her chest was clear to auscultation bilaterally . tachycardia : multifactorial : sepsis , electrolytes abnormalities # . no elevated white blood count but a left shift without bands . her swan-ganz catheter was left in place for hemodynamic monitoring . patient passed spontaneous breathing test on hospital day two and was extubated . sensation was normal bilateral lower and upper extremities . # leukocytosis : patient is afebrile . the remaining paranasal sinuses visualized are clear . external rewarming for hypothermia , check thyroid function . chest x-ray : mild to moderate cardiac enlargement with prominent left ventricle contour . discharge examination : non-focal with normal speech on arrival to the floor , the patient was comfortable and assymptomatic .

# Moving to quantitative comparison: accuracy

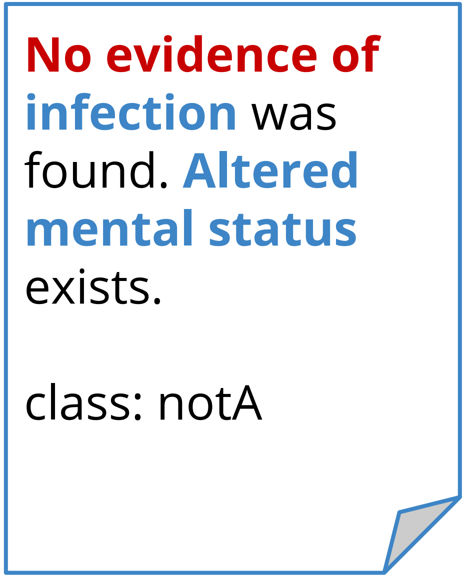
**No evidence of** **infection** was found. **Altered mental status** exists.

class: notA

**Gold Important words:**

1. no
2. evidence
3. of
4. infection
5. altered
6. mental
7. status

# Moving to quantitative comparison: accuracy



**No evidence of infection** was found. **Altered mental status** exists.

class: notA

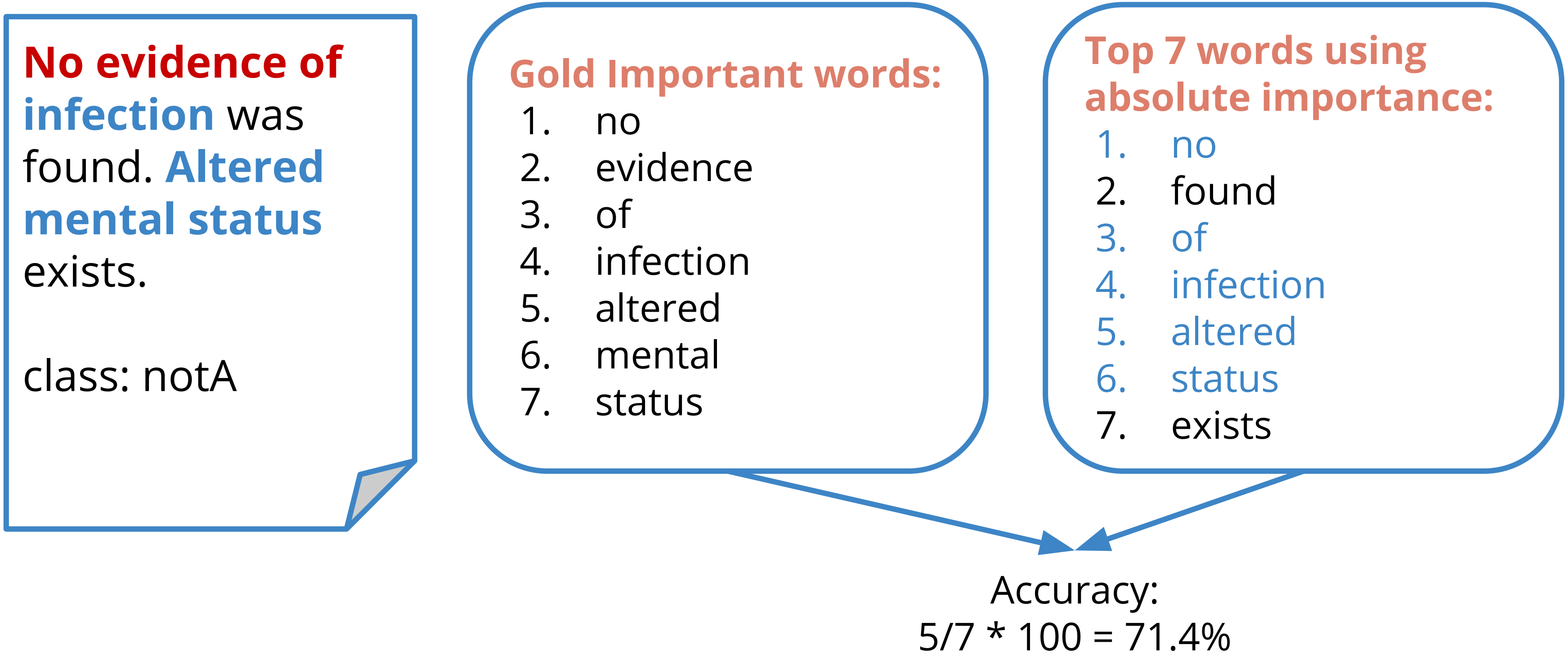
**Gold Important words:**

1. no
2. evidence
3. of
4. infection
5. altered
6. mental
7. status

**Top 7 words using absolute importance:**

1. no
2. found
3. of
4. infection
5. altered
6. status
7. exists

# Moving to quantitative comparison: accuracy



**No evidence of infection** was found. **Altered mental status** exists.

class: notA

**Gold Important words:**

1. no
2. evidence
3. of
4. infection
5. altered
6. mental
7. status

**Top 7 words using absolute importance:**

1. no
2. found
3. of
4. infection
5. altered
6. status
7. exists

Accuracy:
5/7 * 100 = 71.4%

# Mean accuracy(%) of important terms

| Classifier | L2 | sum | dot |
|---|---|---|---|
| LSTM100, E100<br>(0.97) | 17.8 | 13.7 | **26.0** |
| LSTM100, E50<br>(0.96) | 23.7 | 21.5 | **35.4** |
| LSTM50, E100<br>(0.92) | 38.2 | 33.5 | **50.2** |
| LSTM50, E50<br>(0.92) | 26.5 | 25.1 | **36.1** |

# Mean accuracy(%) of important terms

| Classifier | | L2 | sum | dot |
|---|---|---|---|---|
| LSTM100, E100<br>(0.97) | **Why so low?** | 17.8 | 13.7 | 26.0 |
| LSTM100, E50<br>(0.96) | | 23.7 | 21.5 | 35.4 |
| **LSTM50, E100<br>(0.92)** | | **38.2** | **33.5** | **50.2** |
| LSTM50, E50<br>(0.92) | | 26.5 | 25.1 | 36.1 |

# 3. Top skipgrams to explain sequences

Skipgram importance: Mean importance of composed words

document 1 **no** signs **of infection found** topk

document 2 **infection** is **positive**, **found** signs

# 3. Top skipgrams to explain sequences

Skipgram importance: Mean importance of composed words

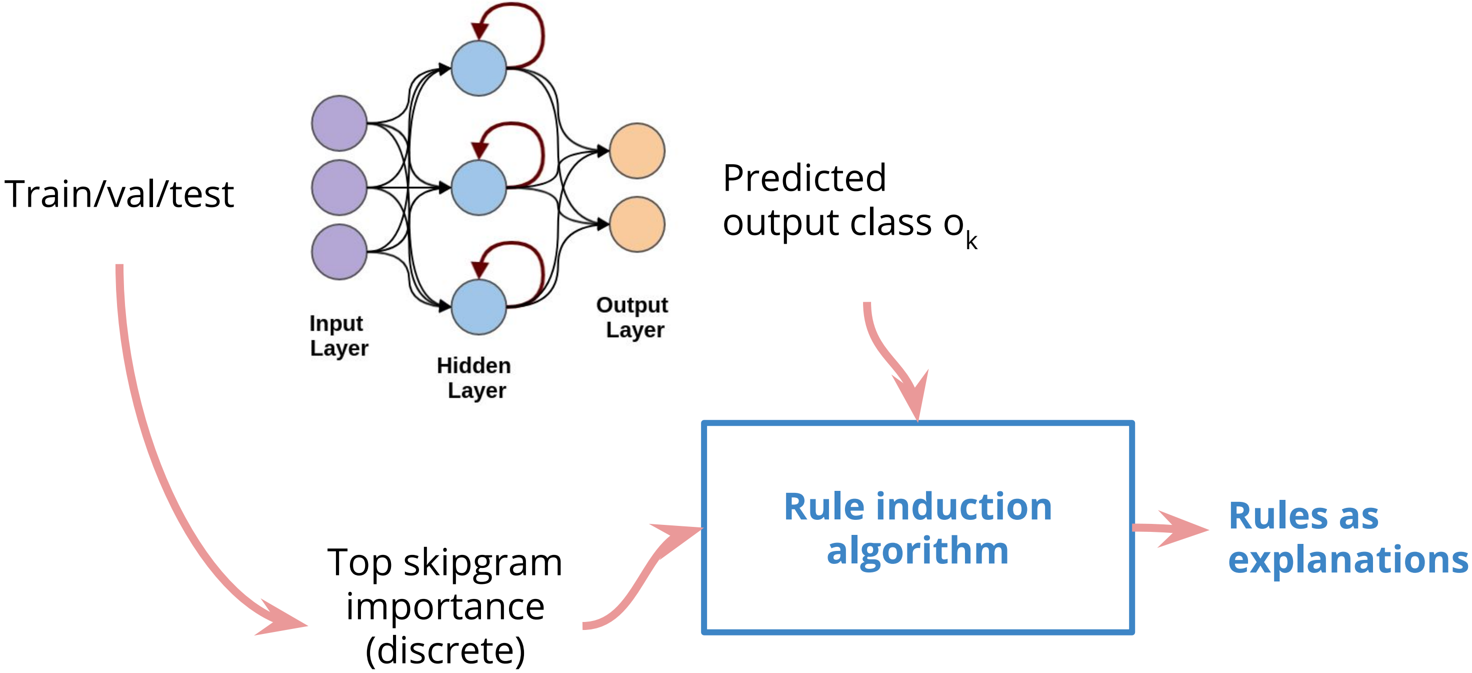
document 1 **no** signs **of infection found**

document 2 **infection** is **positive**, **found** signs

topk

| Most frequent top SGs | **no of infection** | **found** | **infection positive** | |
|---|---|---|---|---|
| document 1 | ✚ | ▬ | 🚫 | class A |
| document 2 | 🚫 | ✚ | ▬ | class notA |

# 5. Inducing rules as explanations



Train/val/test

Predicted output class $o_k$

Top skipgram importance (discrete)

**Rule induction algorithm**

**Rules as explanations**

# Results

Explanation fidelity (test): 0.9+ macro F-score

Example rules:

***infection = neg*** ⇒ ***non_septic*** (✓ 2899/2985)

***infection = pos*** AND ***tachypnea = pos*** ⇒ ***septic*** (✓ 1263/1285)

***urinary tract = absent*** AND ***source infection = absent*** AND ***meningitis = pos***
AND ***tachypnea = absent*** ⇒ ***septic*** (✓ 1474/1556)

# Observations

Best performing LSTM:

- Low coverage score of important terms compared to gold (only 26%)

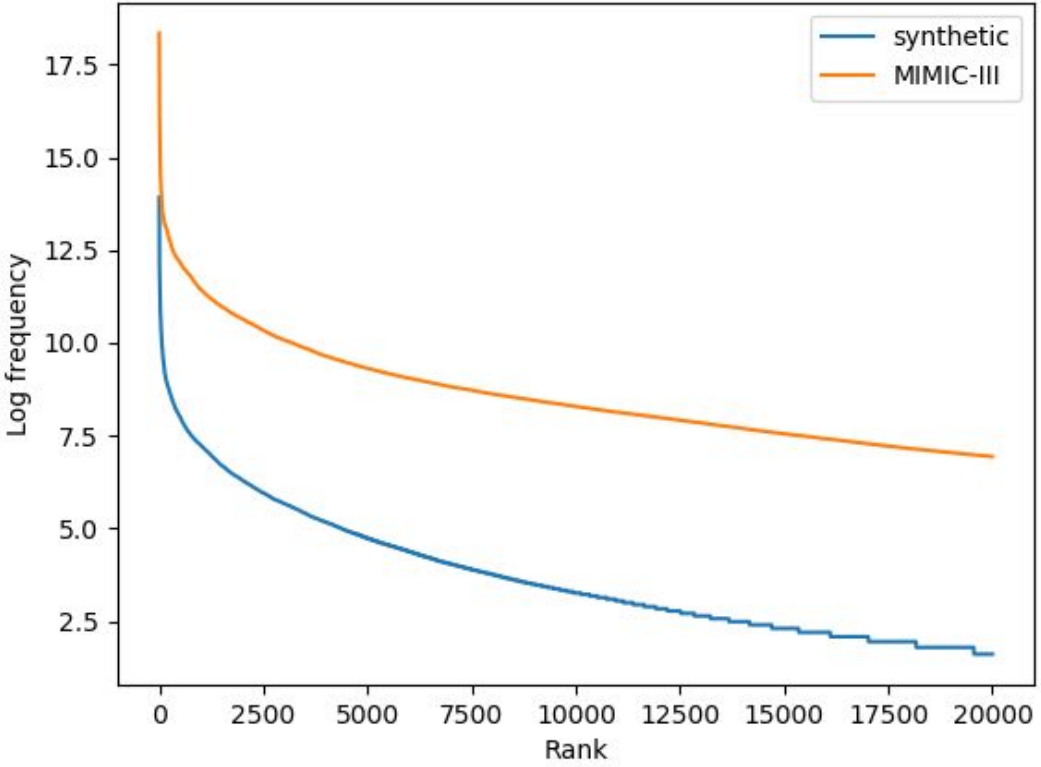
- Has more diffused smaller gradient values

- Has more conditions in rule explanations (consistently)

**Is the model more generalized?**

Question by Elves Sousa from the Noun Project

# Zipfian distribution: MIMIC-III vs synthetic


synthetic
MIMIC-III
Log frequency
Rank