Enhancing Human-in-the-Loop Learning for Binary Sentiment Word Classification

Belén Martín-Urcelay, Christopher R. Rozell, Matthieu R. Bloch

December 17, 2024

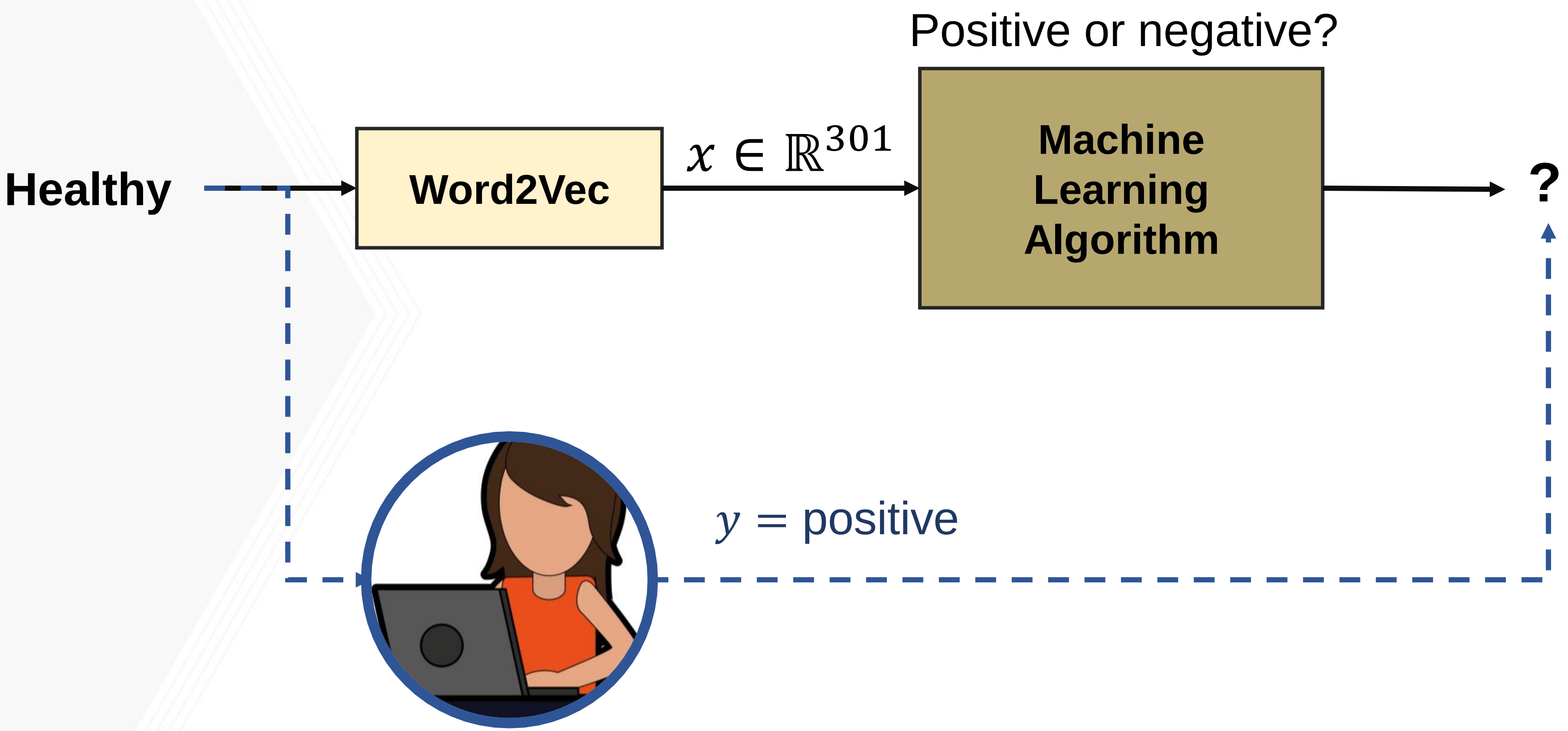
Georgia Tech

CREATING THE NEXT

CDC·2024 MILANO

1. Background Motivation

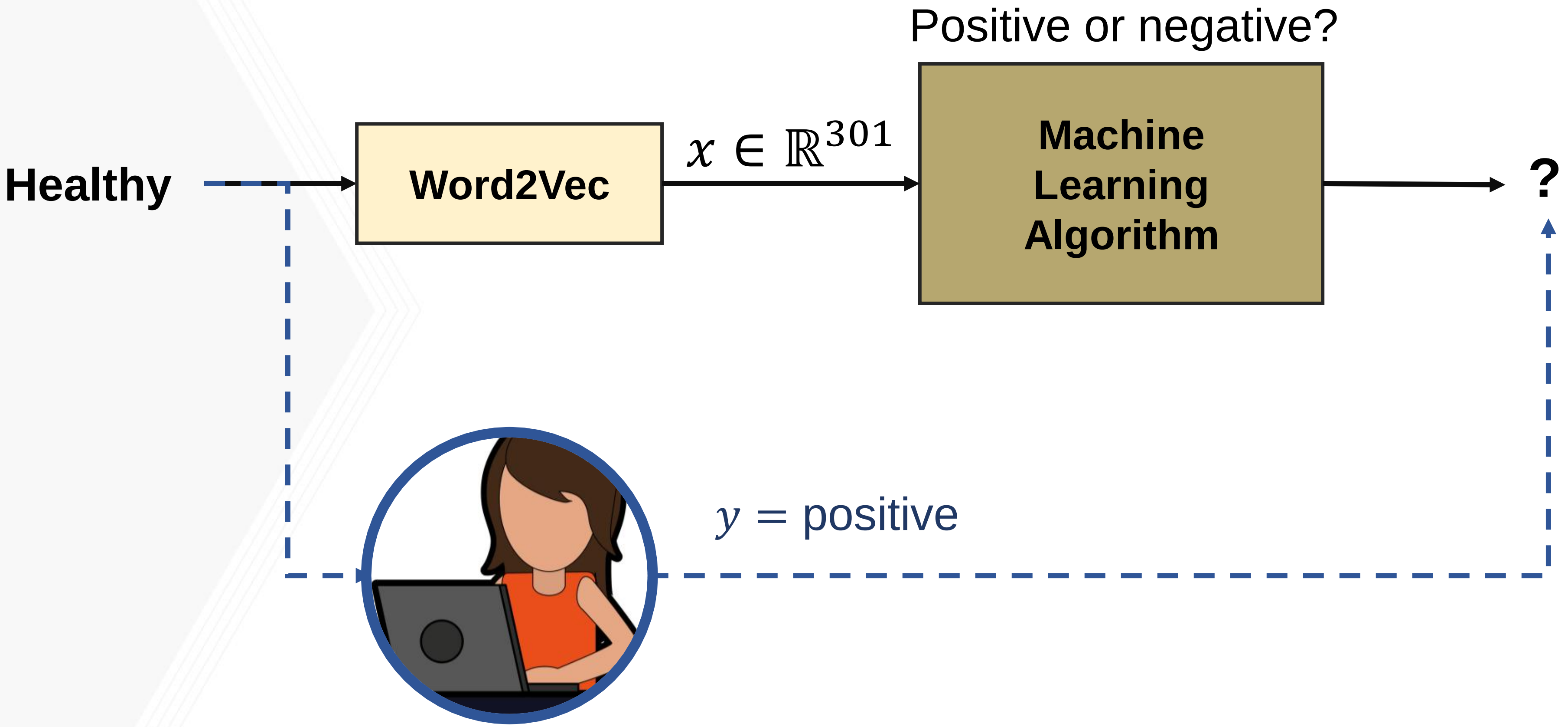
How can ML algorithms learn to classify words **without exhausting human patience?**



Supervised learning: training with labels from humans

1. Background Motivation

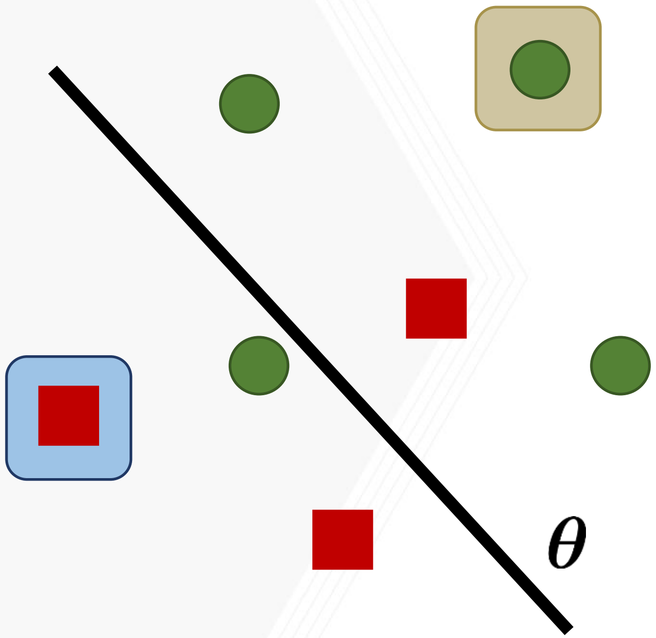
How can ML algorithms learn to classify words without exhausting human patience?



- ❑ Could we get **more information** from the human?

2. More Information

Queries must be human and mathematically interpretable



● :	Positive word embedding
■ :	Negative word embedding
θ :	Word Classifier

"Select the example that you find…"

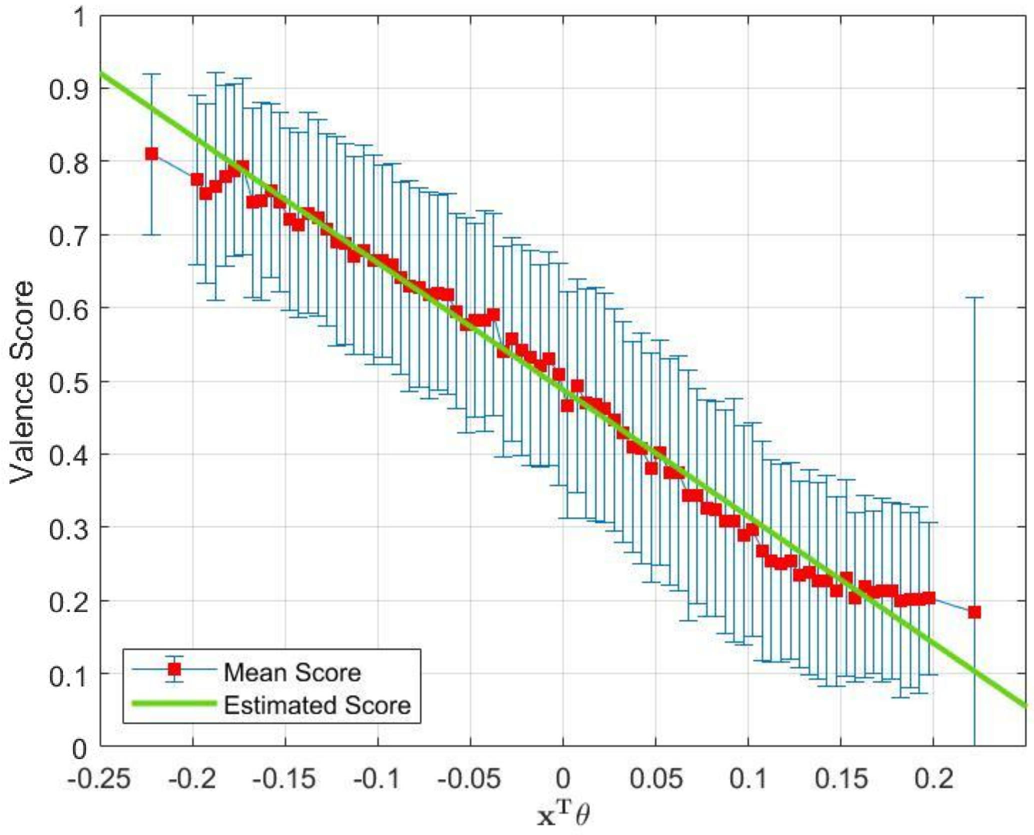
- Query: "… most **positive**"
- Query: "… most **negative**"

Hypothesis: Human answers depend on the distance to the ground truth

Georgia Tech
CREATING THE NEXT

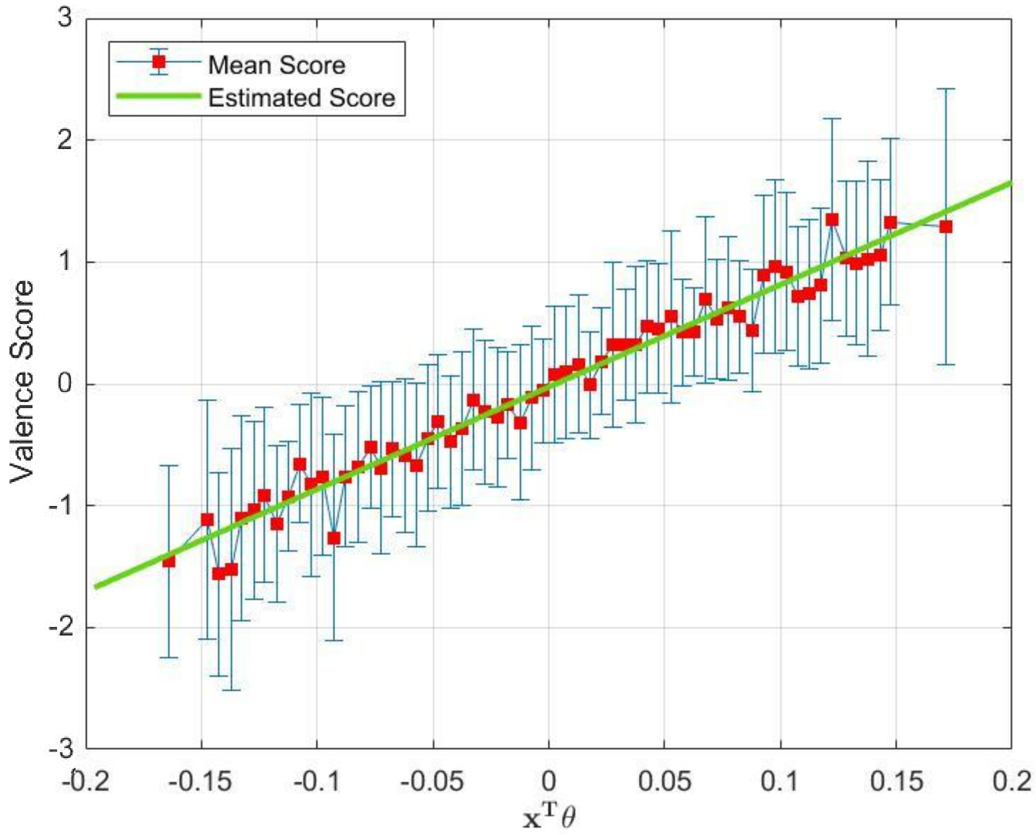
2. More Information Valence vs Distance to θ

NRC-VAD Lexicon



$$\mathbb{E}\left[\text{score}(\mathbf{x})|\boldsymbol{\theta}\right] = -1.73(\mathbf{x}^T\boldsymbol{\theta}) + 0.49$$

SocialNet



$$\mathbb{E}\left[\text{score}(\mathbf{x})|\boldsymbol{\theta}\right] = 6.96(\mathbf{x}^T\boldsymbol{\theta}) - 0.09$$

$\mathbf{x}$: Word embedding $\boldsymbol{\theta}$: Ground truth classifier

2. More Information Multinomial Logit Model

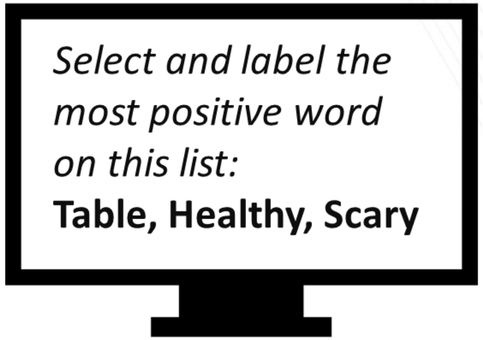
Queries must be human and mathematically interpretable

“Select and label the most positive/negative word”

$$\mathbb{P}\left[y=1|\mathbf{x}\right] = \frac{1}{1+\exp\left(W(\boldsymbol{\theta}^T\mathbf{x})\right)}, \text{ with } W \in \mathbb{R}.$$

$$\mathbb{P}\left[\mathbf{x}_i|\{\mathbf{x}_j\}_{j=1}^N, \boldsymbol{\theta}\right] = \frac{\exp\left(u(\mathbf{x}_i,\boldsymbol{\theta})\right)}{\sum_{j=1}^N \exp\left(u(\mathbf{x}_j,\boldsymbol{\theta})\right)}$$

Likelihood



- Q1: Most positive

$$u(\mathbf{x},\boldsymbol{\theta}) = \frac{a}{\sigma}\mathbf{x}^T\boldsymbol{\theta}$$

- Q2: Most negative

$$u(\mathbf{x},\boldsymbol{\theta}) = -\frac{a}{\sigma}\mathbf{x}^T\boldsymbol{\theta}$$

$\mathbf{x}$: Word embedding

$\boldsymbol{\theta}$: Ground truth classifier

y: Word label (positive: 1, negative: 0)

u: Utility function (expected valence score)

2. More Information Multinomial Logit Model

$$\mathbb{P}\left[y=1|\mathbf{x}\right] = \frac{1}{1+\exp\left(W(\boldsymbol{\theta}^T\mathbf{x})\right)}, \text{ with } W \in \mathbb{R}.$$

$$\mathbb{P}\left[\mathbf{x}_i|\{\mathbf{x}_j\}_{j=1}^N, \boldsymbol{\theta}\right] = \frac{\exp\left(u(\mathbf{x}_i,\boldsymbol{\theta})\right)}{\sum_{j=1}^N \exp\left(u(\mathbf{x}_j,\boldsymbol{\theta})\right)}$$

Likelihood

How can we get the posterior?

$$p_{\boldsymbol{\theta}} = \mathbb{P}\left[\boldsymbol{\theta}|\{\mathbf{x}_t, y_t, q_t, \{\mathbf{x}_j\}_{j=1}^N\}_{t=0}^i\right]$$

→ Approximate $\boldsymbol{\theta}$ as a multivariate Gaussian

Posterior over θ

$\mathbf{x}$: Word embedding

$\boldsymbol{\theta}$: Ground truth classifier

q: Query

y: Word label

u: Utility function (expected valence score)

2. More Information Update Given Label

$$\mathbb{P}\left[y=1|\mathbf{x}\right] = \frac{1}{1+\exp\left(W(\boldsymbol{\theta}^T\mathbf{x})\right)}, \text{ with } W \in \mathbb{R}.$$

How do we update the posterior given the label?

Jaakkola and Jordan give a closed form approximation*

$$\begin{cases} \boldsymbol{\Sigma}_{\text{pos}}^{-1} = \boldsymbol{\Sigma}^{-1} + 2\frac{\tanh(\xi/2)}{4\xi}W^2\mathbf{x}_i\mathbf{x}_i^T \\ \boldsymbol{\mu}_{\text{pos}} = \boldsymbol{\Sigma}_{\text{pos}}\left[\boldsymbol{\Sigma}^{-1}\mu + \left(y_i - \frac{1}{2}\right)W\mathbf{x}_i\right] \\ \xi^2 = W^2\mathbf{x}_i^T\boldsymbol{\Sigma}_{\text{pos}}\mathbf{x}_i + W^2(\mathbf{x}_i^T\boldsymbol{\mu}_{\text{pos}})^2 \end{cases}$$

Bayesian parameter estimation via variational methods – Jaakkola and Jordan, 2000

$\mathbf{x}$: Word embedding

y: Word label

Prior: $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Posterior: $\mathcal{N}(\boldsymbol{\mu}_{\text{pos}}, \boldsymbol{\Sigma}_{\text{pos}})$

$\boldsymbol{\theta}$: Ground truth classifier

2. More Information Update Given Word

$$\mathbb{P}\left[\mathbf{x}_i \middle| \{\mathbf{x}_j\}_{j=1}^N, \boldsymbol{\theta}\right] = \frac{\exp\left(K\mathbf{x}_i^T\boldsymbol{\theta}\right)}{\sum_{j=1}^N \exp\left(K\mathbf{x}_j^T\boldsymbol{\theta}\right)}$$

How do we update the posterior given the word selected?

$$ELBO(q) = -\mathrm{KL}(q(\boldsymbol{\theta})\|p(\boldsymbol{\theta})) + \mathbb{E}_{\boldsymbol{\theta}\sim q}\left[K\mathbf{x}_s^T\boldsymbol{\theta}\right] - \mathbb{E}_{\boldsymbol{\theta}\sim q}\left[\log\sum_{j=1}^{|\mathcal{S}|}\exp\left(K\mathbf{x}_j^T\boldsymbol{\theta}\right)\right]$$

$$KL(q||p) = \frac{1}{2}\left[\log\frac{|\Sigma_p|}{|\Sigma_q|} - d + (\boldsymbol{\mu_q})^T\Sigma_p^{-1}(\boldsymbol{\mu_q}) + (\boldsymbol{\mu_p})^T\Sigma_p^{-1}(\boldsymbol{\mu_p}) - 2(\boldsymbol{\mu_q})^T\Sigma_p^{-1}(\boldsymbol{\mu_p}) + tr\left\{\Sigma_p^{-1}\Sigma_q\right\}\right]$$

$$\mathbb{E}_{\boldsymbol{\theta}\sim q}\left[K\mathbf{x}_s^T\boldsymbol{\theta}\right] = K\mathbf{x}_s^T\boldsymbol{\mu}_q$$

$$\mathbb{E}_{\boldsymbol{\theta}\sim q}\left[\log\sum_{j=1}^{|\mathcal{S}|}\exp\left(K\mathbf{x}_j^T\boldsymbol{\theta}\right)\right] \geq \log\sum_{j=1}^{|\mathcal{S}|}\exp\left(K\mathbf{x}_j^T\boldsymbol{\mu}_q + 0.5\mathbf{x}_j^T\boldsymbol{\Sigma}_q\mathbf{x}_j\right)$$ *[Braun and McAuliffe, 2007]*

$p(\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$: Prior $\qquad$ $q(\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$: Variational Distribution

2. More Information Multinomial Logit Model

$$\mathbb{P}\left[\mathbf{x}_i | \{\mathbf{x}_j\}_{j=1}^{N}, \boldsymbol{\theta}\right] = \frac{\exp\left(u(\mathbf{x}_i, \boldsymbol{\theta})\right)}{\sum_{j=1}^{N} \exp\left(u(\mathbf{x}_j, \boldsymbol{\theta})\right)}$$

$$\mathbb{P}\left[y = 1 | \mathbf{x}\right] = \frac{1}{1+\exp\left(W(\boldsymbol{\theta}^T \mathbf{x})\right)}, \text{ with } W \in \mathbb{R}.$$

Likelihood

Variational inference

Posterior over θ

$\mathbf{x}$: Word embedding

$\boldsymbol{\theta}$: Ground truth classifier

y: Word label (positive: 1, negative: 0)

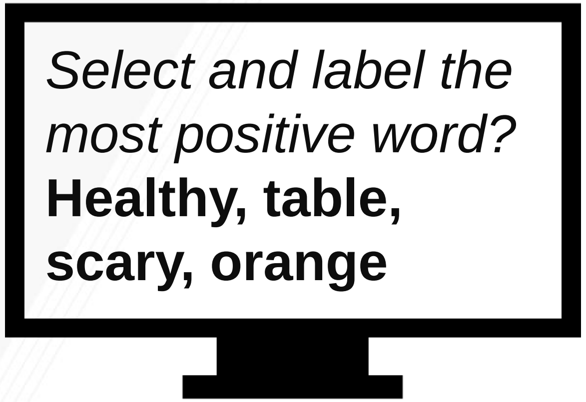
u: Utility function (expected valence score)

2. More Information Experiments

Label:

$s_{\text{healthy}} > 0.5$

Label + Word:



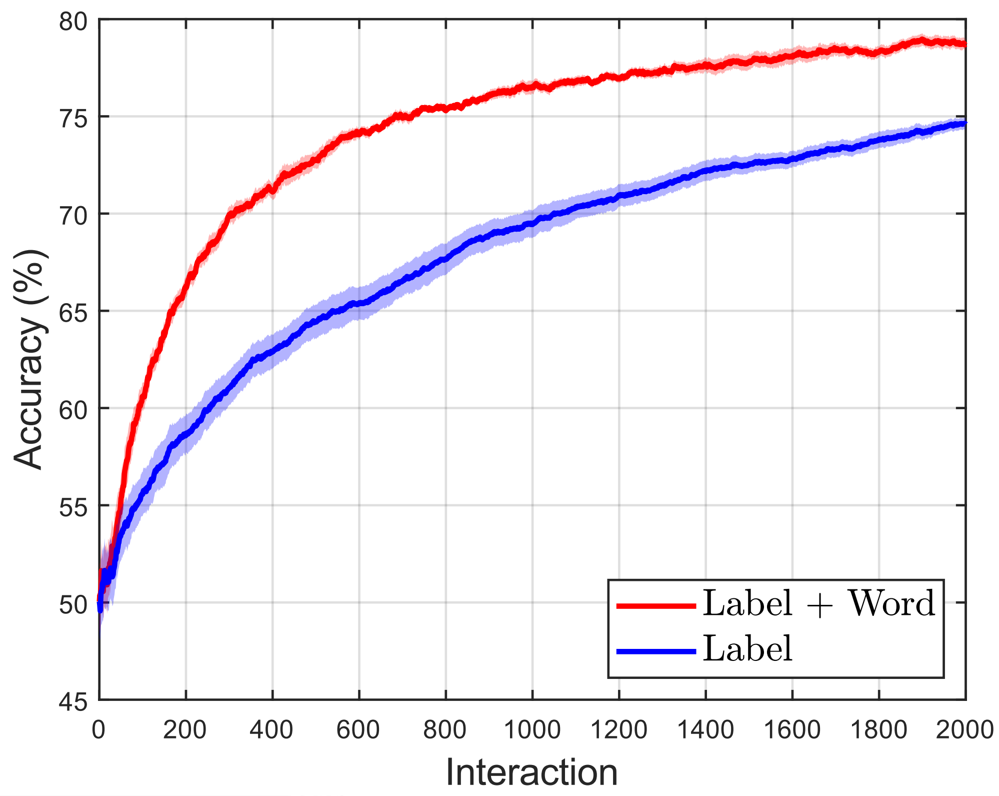
$s_{\text{healthy}} > s_{\text{table}}, s_{\text{scary}}, s_{\text{orange}}$

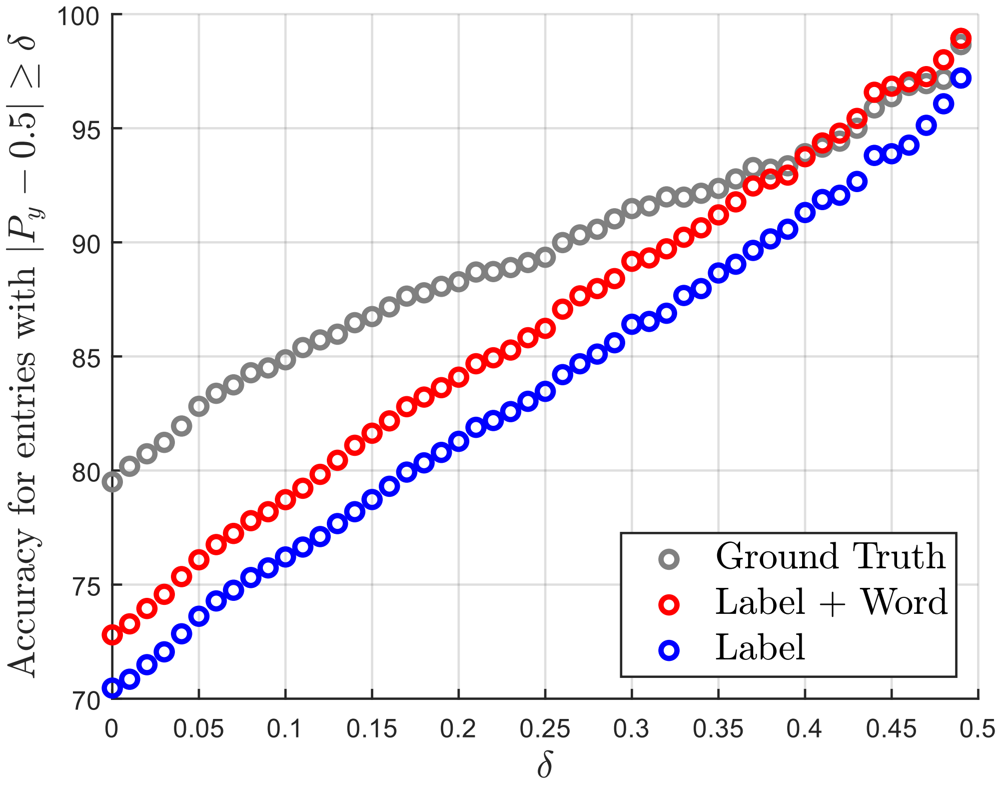
$s_{\text{healthy}} > 0.5$

NRC-VAD Lexicon dataset:

Each word $\mathbf{x}$ has a valence score mean $\mu_{\mathbf{x}}$ and variance $\sigma^2_{\mathbf{x}} \rightarrow \text{score: } s_{\mathbf{x}} \sim \mathcal{N}(\mu_{\mathbf{x}}, \sigma^2_{\mathbf{x}})$

2. More Information Empirical Results





2. More Information Theoretical Results

Under Assumptions

1. Word and label selected are independent of the history given the classifier

   $p(\mathbf{x}_t, y_t|\boldsymbol{\theta}, q_t, \mathcal{S}_t, \mathcal{F}_{t-1}) = p(\mathbf{x}_t, y_t|\boldsymbol{\theta}, q_t, \mathcal{S}_t)$

2. The label only depends on the word it is referring to

   $p(y_t|\mathbf{x}_t, q_t, \mathcal{S}_t) = p(y_t|\mathbf{x}_t)$

3. An answer always provides some information

   $I(\boldsymbol{\theta}; X_t, Y_t|\mathcal{F}_{t-1}) \geq L > 0$

Simplified Theorem

The expected stopping time $T_\epsilon = \min\{t : \left|\boldsymbol{\Sigma}_{\boldsymbol{\theta}|\mathcal{F}_t}\right|^{1/d} < \epsilon\}$ is bounded as

$$\frac{d}{2}\frac{\log_2 \frac{2}{\pi e \epsilon}}{\log_2 2|\mathcal{S}|} \leq \mathbb{E}[T_\epsilon] \leq \frac{d}{2L}\log_2 \frac{e^4 d^2}{2\sqrt{2}(d+2)\epsilon} - 1.$$

where $\mathbf{x} \in \mathbb{R}^d$ and $\mathcal{S} = \{\mathbf{x}_j\}_{j=1}^{|\mathcal{S}|}$ are the candidate words.

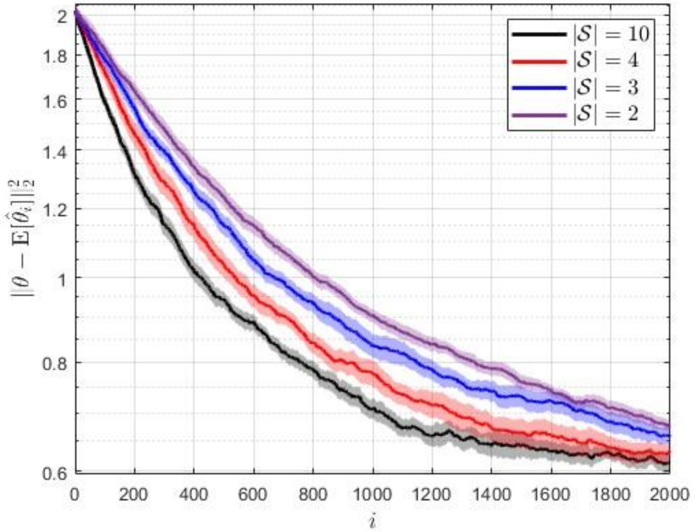
2. More Information Theoretical Results

Simplified Theorem

The expected stopping time $T_\epsilon = \min\{t : \left|\boldsymbol{\Sigma}_{\boldsymbol{\theta}|\mathcal{F}_t}\right|^{1/d} < \epsilon\}$ is bounded as

$$\frac{d}{2}\frac{\log_2 \frac{2}{\pi e \epsilon}}{\log_2 2|\mathcal{S}|} \leq \mathbb{E}[T_\epsilon] \leq \frac{d}{2L}\log_2 \frac{e^4 d^2}{2\sqrt{2}(d+2)\epsilon} - 1.$$

where $\mathbf{x} \in \mathbb{R}^d$ and $\mathcal{S} = \{\mathbf{x}_j\}_{j=1}^{|\mathcal{S}|}$ are the candidate words.

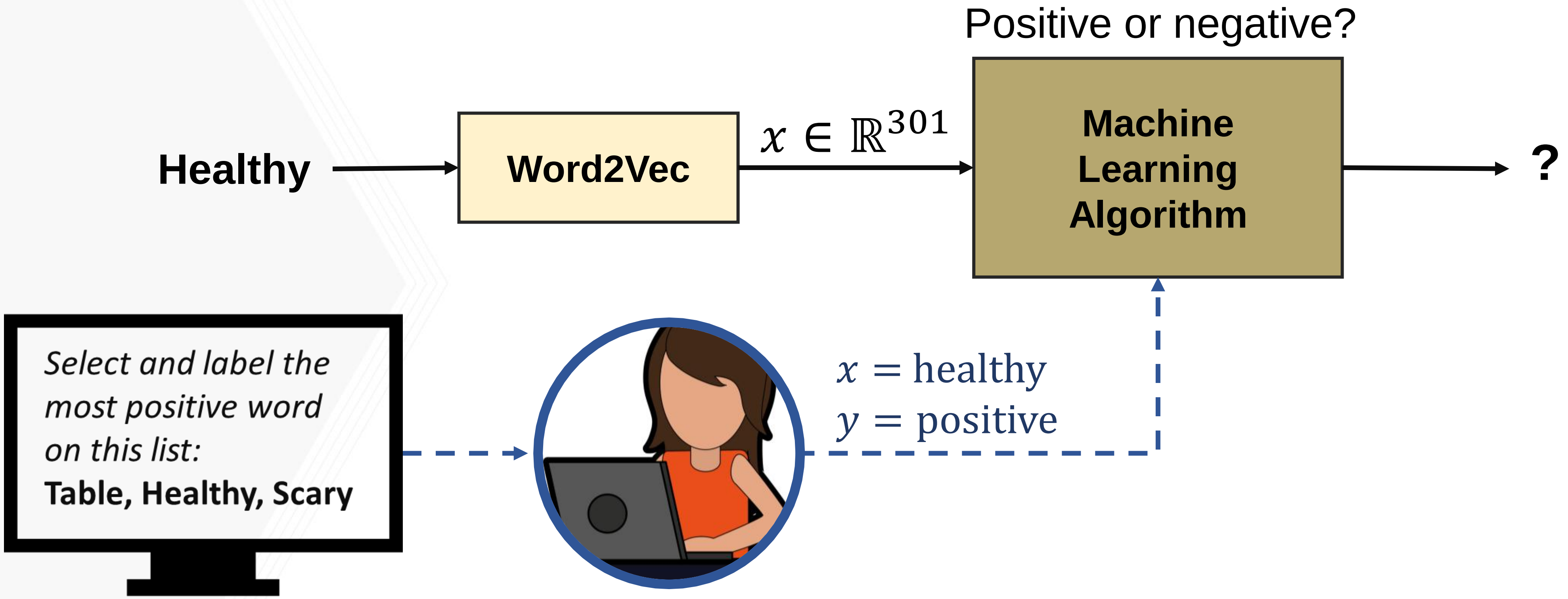


- ✓ The number of question to ask the human to reach uncertainty $< \epsilon$ is on the order of $\log 1/\epsilon$
- ✓ Related to the error $\mathrm{MSE}_t = \mathrm{trace}(\Sigma_{\boldsymbol{\theta}|\mathcal{F}_t}) \geq d|\Sigma_{\boldsymbol{\theta}|\mathcal{F}_t}|^{1/d}$
- ✓ The more words in the list, the faster the error decays

Georgia Tech
CREATING THE NEXT

2. More Information Motivation

How can ML algorithms learn to classify words without exhausting human patience?

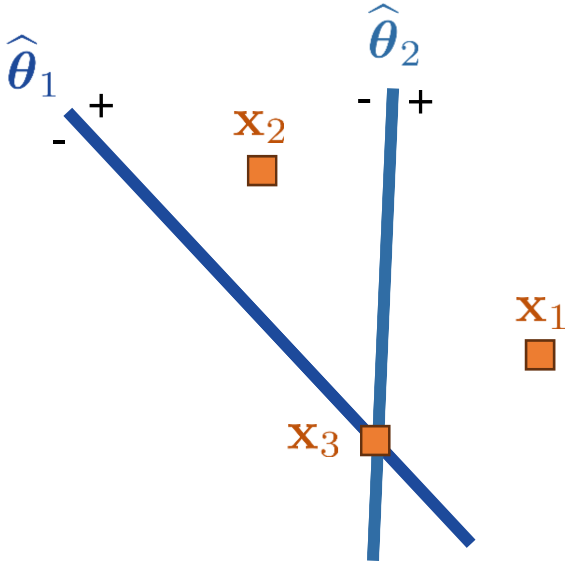


- ✓ Could we get **more information** from the human?
- ❑ Could we get the information **faster**?

3. Faster Active Learning Heuristic

Instead of showing the humans a random list of words, could we select them in a smart way?

Example	$P[+\|x, \theta = 1]$	$P[+\|x, \theta = 2]$
x_1	0.99	0.99
x_2	0.99	0.01
x_3	0.5	0.5



Heuristic in AL:

$$\text{argmax}_{\mathbf{x}} \underbrace{H\left(\mathbb{E}_{\boldsymbol{\theta}}\left[f_\theta(\mathbf{x})\right]\right)} - \underbrace{\mathbb{E}_\theta\left[H\left(f_\theta(\mathbf{x})\right)\right]}$$

→ Maximize uncertainty of the expected output

→ Minimize uncertainty due to noise

H: Entropy $\boldsymbol{\theta}$: Ground truth classifier $\mathbf{x}$: Word embedding

3. Faster Active Word Selection

Heuristic:

$$\mathcal{S} = \text{argmax}_{\mathcal{S} \in \{\mathcal{X}\}^k} \underbrace{H\left(\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbf{x}_i, y_i \mid q, \boldsymbol{\theta}, \mathcal{S}\right]\right)} - \underbrace{\mathbb{E}_{\boldsymbol{\theta}}\left[H\left(\mathbf{x}_i, y_i \mid q, \boldsymbol{\theta}, \mathcal{S}\right)\right]}$$

→ Maximize uncertainty of the expected output

→ Minimize uncertainty due to noise

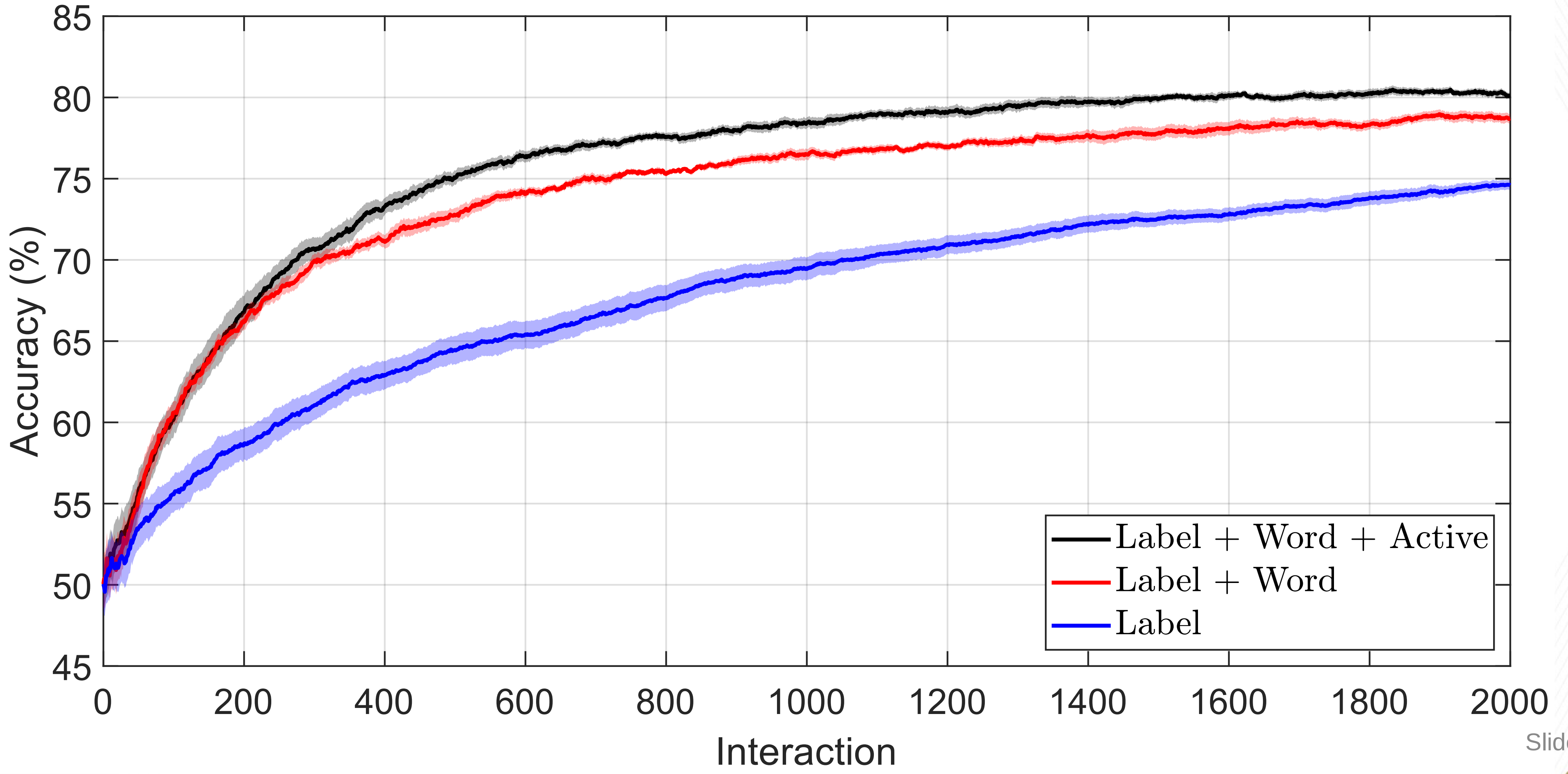
Problem: There exist combinatorically many sets $\binom{|\mathcal{X}|}{|\mathcal{S}|}$ to maximize over

If $|\mathcal{X}| = 3500$ and $|\mathcal{S}| = 4$:
$\binom{|\mathcal{X}|}{|\mathcal{S}|} \sim 10^{15}$
Galaxies $\sim 10^{12}$
If each computation 1ms,
$t \sim 31,7000$ years

Our Approach: Greedily select one word at a time

H: Entropy $\boldsymbol{\theta}$: Ground truth classifier $\mathbf{x}$: Word embedding y: Word label

3. Faster Results



4. Summary

1. We introduce a novel **human response model**.
2. We **speed up learning** in sentiment classification
   - By combining label requests with word selection.
   - By active query selection.
3. We **validate** our approach
   - Theoretically: Bounds for expected stopping time.
   - Empirically: Experiments with human data.

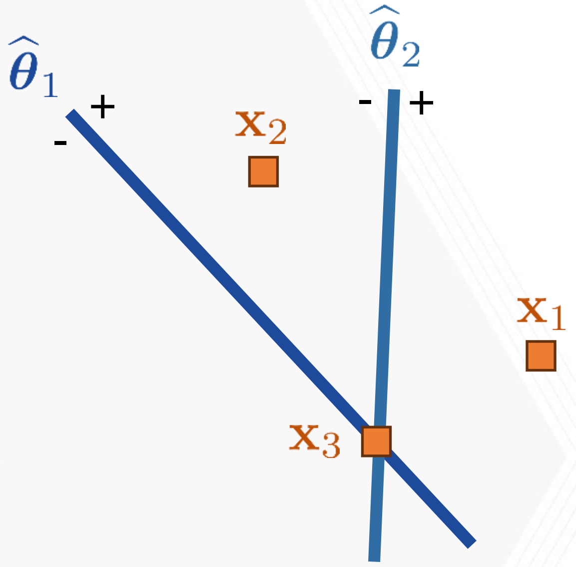
Enhancing Human-in-the-Loop Learning for Binary Sentiment Word Classification

Belén Martín-Urcelay, Christopher R. Rozell, Matthieu R. Bloch

December 17, 2024

CDC·2024
MILANO

3. Faster Active Learning Heuristic



Example	$P[+\|x, \theta = 1]$	$P[+\|x, \theta = 2]$
x_1	0.99	0.99
x_2	0.99	0.01
x_3	0.5	0.5

$x_1 \rightarrow H(0.99) - [0.5 (H(0.99) + H(0.99))] = 0.02 - 0.02 = 0$

$x_2 \rightarrow H(0.5) - [0.5 (H(0.01) + H(0.99))] = 1 - 0.02 = 0.98$

$x_3 \rightarrow H(0.5) - [0.5 (H(0.5) + H(0.5))] = 1 - 1 = 0$

Heuristic in AL:

$$\text{argmax}_{\mathbf{x}} \underbrace{H\left(\mathbb{E}_{\boldsymbol{\theta}}\left[f_\theta(\mathbf{x})\right]\right)} - \underbrace{\mathbb{E}_\theta\left[H\left(f_\theta(\mathbf{x})\right)\right]}$$

→ Maximize uncertainty of the expected output

→ Minimize uncertainty due to noise

H: Entropy $\boldsymbol{\theta}$: Ground truth classifier $\mathbf{x}$: Word embedding