Learning Only from Relevant Keywords and Unlabeled Documents

Nontawat Charoenphakdee[1,3] Jongyeong Lee[1,3]
Yiping Jin[2] Dittaya Wanvarie[2] Masashi Sugiyama[3,1]
1: The University of Tokyo 2: Chulalongkorn University 3:RIKEN AIP

Summary

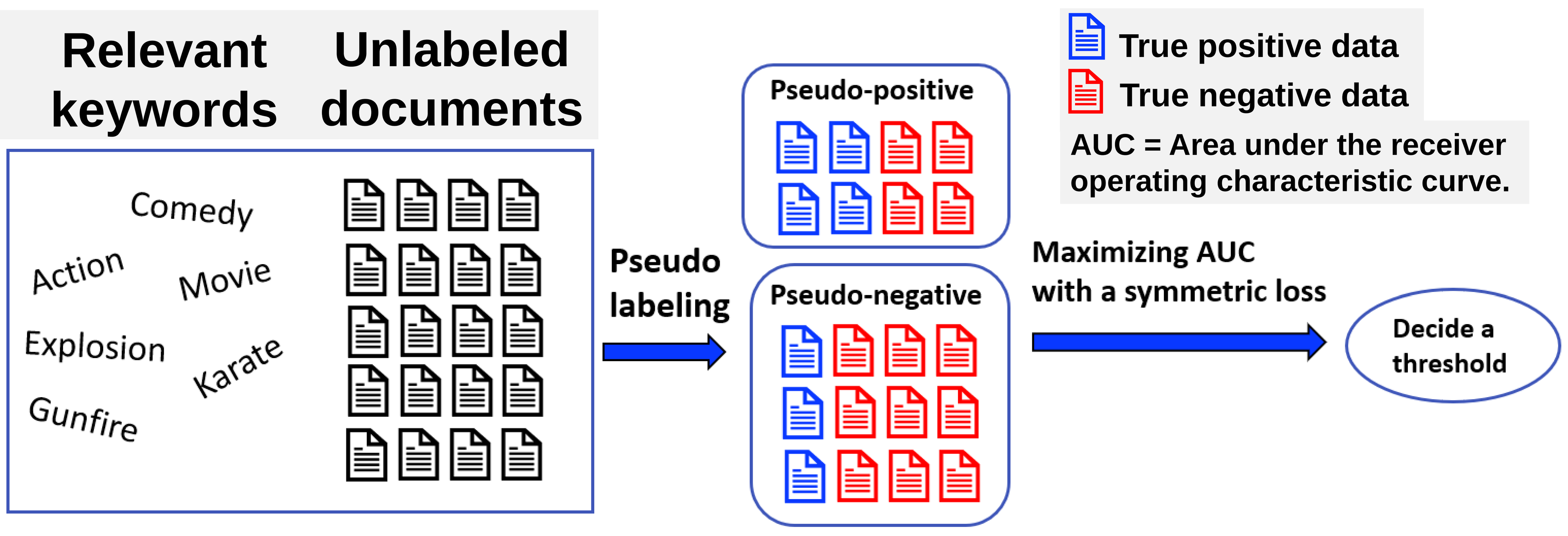


Figure 1: Example where a target class is "Action movie".

Common approach: Liu+ (2004); Druck+ (2008); Settle (2011); Jin+, (2017)

1. Use keywords to pseudo-label the unlabeled data
2. Learn a classifier from pseudo-labeled data

But pseudo-labeling can be unreliable!

- True positive can be pseudo-labeled as negative (and vice versa)
- Theoretical understanding of this problem is limited

Proposed: maximize AUC then find a threshold.

- AUC can be maximized even pseudo-labeling is imperfect
- Theoretically guaranteed with estimation error bound
- Also allows flexible choices of model and optimization algorithm

AUC maximization from pseudo-labeled data

Given: Two sets of documents

Pseudo-positive:

$$\{\boldsymbol{x}_i^{\mathrm{CP}}\}_{i=1}^{n_{\mathrm{CP}}} \overset{\text{i.i.d.}}{\sim} \theta \mathrm{pos}(\boldsymbol{x}) + (1-\theta)\,\mathrm{neg}(\boldsymbol{x})$$

$\mathrm{pos}(\boldsymbol{x})$: $p(\boldsymbol{x}|y=+1)$
$\mathrm{neg}(\boldsymbol{x})$: $p(\boldsymbol{x}|y=-1)$
$\theta, \theta' \in [0,1]$ and $\theta > \theta'$

Pseudo-negative:

$$\{\boldsymbol{x}_j^{\mathrm{CN}}\}_{j=1}^{n_{\mathrm{CN}}} \overset{\text{i.i.d.}}{\sim} \theta' \mathrm{pos}(\boldsymbol{x}) + (1-\theta')\,\mathrm{neg}(\boldsymbol{x})$$

$\mathbb{E}_{\mathrm{P}}[\cdot] : \underset{\boldsymbol{x}\sim\mathrm{pos}(\boldsymbol{x})}{\mathbb{E}}[\cdot]$
$\mathbb{E}_{\mathrm{N}}[\cdot] : \underset{\boldsymbol{x}\sim\mathrm{neg}(\boldsymbol{x})}{\mathbb{E}}[\cdot]$

Find: $g: \mathcal{X} \to \mathbb{R}$ that **minimizes** AUC risk

$$R_{\mathrm{AUC}}^{\ell_{0\text{-}1}}(g) = \mathbb{E}_{\mathrm{P}}[\mathbb{E}_{\mathrm{N}}[\ell_{0\text{-}1}(g(\boldsymbol{x}^{\mathrm{P}}) - g(\boldsymbol{x}^{\mathrm{N}}))]]$$

Zero-one loss

$$\ell_{0\text{-}1}(z) = \begin{cases} 0, z > 0 \\ \frac{1}{2}, z = 0 \\ 1, z < 0 \end{cases}$$

AUC risk is with respect to the clean data.

How to minimize the clean risk using pseudo-labeled data?

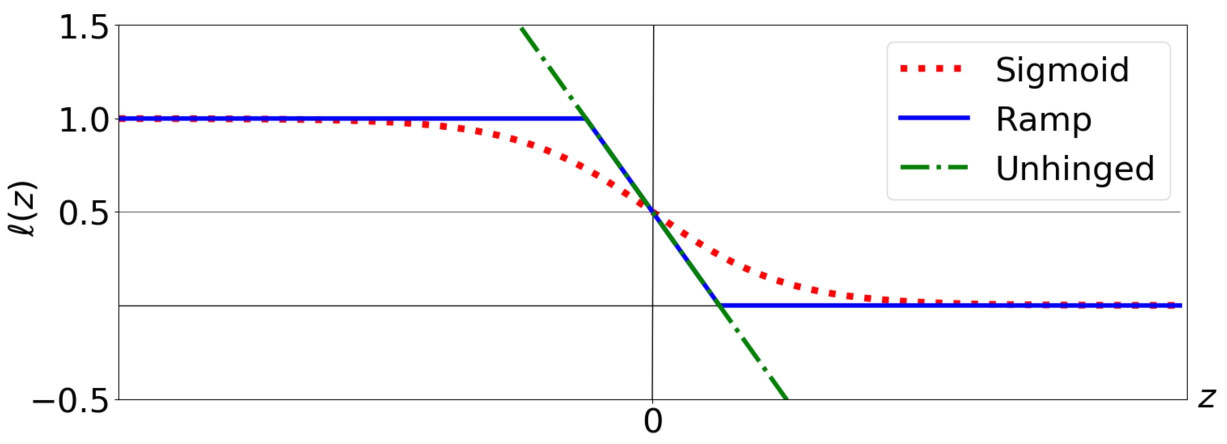
Relationship between **pseudo-labeled** and **clean** risks:

For any loss $\ell: \mathbb{R} \to \mathbb{R}$, $\varphi^{\ell}(\boldsymbol{x}, \boldsymbol{x}') = \ell(g(\boldsymbol{x}) - g(\boldsymbol{x}')) + \ell(g(\boldsymbol{x}') - g(\boldsymbol{x}))$

$$\underbrace{R_{\mathrm{AUC\text{-}Corr}}^{\ell}(g)}_{\text{Pseudo-labeled risk}} = (\theta - \theta')\underbrace{R_{\mathrm{AUC}}^{\ell}(g)}_{\text{Clean risk}} + \underbrace{(1-\theta)\theta'\mathbb{E}_{\mathrm{P}}[\mathbb{E}_{\mathrm{N}}[\varphi^{\ell}(\boldsymbol{x}^{\mathrm{P}}, \boldsymbol{x}^{\mathrm{N}})]]}_{\text{Excessive term}}$$
$$+ \underbrace{\frac{\theta\theta'}{2}\mathbb{E}_{\mathrm{P}'}[\mathbb{E}_{\mathrm{P}}[\varphi^{\ell}(\boldsymbol{x}^{\mathrm{P}'}, \boldsymbol{x}^{\mathrm{P}})]]}_{\text{Excessive term}} + \underbrace{\frac{(1-\theta)(1-\theta')}{2}\mathbb{E}_{\mathrm{N}'}[\mathbb{E}_{\mathrm{N}}[\varphi^{\ell}(\boldsymbol{x}^{\mathrm{N}'}, \boldsymbol{x}^{\mathrm{N}})]]}_{\text{Excessive term}}.$$

Minimizing pseudo-labeled risk suffers from excessive terms.

Using symmetric loss: $\ell^{\mathrm{sym}}(z) + \ell^{\mathrm{sym}}(-z) = K$



K : constant

$$\underbrace{R_{\mathrm{AUC\text{-}Corr}}^{\ell_{\mathrm{sym}}}(g)}_{\text{Pseudo-labeled risk}} = (\theta - \theta')\underbrace{R_{\mathrm{AUC}}^{\ell_{\mathrm{sym}}}(g)}_{\text{Clean risk}} + \underbrace{\frac{K(1-\theta+\theta')}{2}}_{\text{Constant}},$$

the minimizers of both risks are identical!

Charoenphakdee+ (2019)

Theoretical analysis

Estimation error bound:

$$\underbrace{R_{\mathrm{AUC}}^{\ell_{\mathrm{sym}}}(\hat{g}) - R_{\mathrm{AUC}}^{\ell_{\mathrm{sym}}}(g^*)}_{\text{Clean risk}} \leq \underbrace{\frac{1}{\theta - \theta'}}_{\text{Pseudo-labeling quality}}\left[\mathcal{O}_p\left(\underbrace{\sqrt{\frac{1}{n_{\mathrm{CP}}} + \frac{1}{n_{\mathrm{CN}}}}}_{\text{Converge to zero as data size increases}}\right)\right]$$

$\hat{g}$: Function learned from our framework
g^*: True minimizer

Estimation error converges to zero as $n_{\mathrm{CP}}, n_{\mathrm{CN}} \to \infty$.

$\hat{g}$ converges to g^* as the number of data increases!

Threshold selection

A reasonable threshold $\beta_{\mathrm{BEP}} \in \mathbb{R}$ can be obtained if positive data ratio π of unlabeled documents is known.

$$\pi \approx \frac{1}{n}\sum_{\boldsymbol{x}\in \mathrm{D}} \mathrm{sign}(g(\boldsymbol{x}) - \beta_{\mathrm{BEP}})$$

n: Training data size
D: Training documents

Known as **precision-recall breakeven point (BEP).** Kato+ (2019)

We can re-adjust a bad threshold caused by pseudo-labeling using β_{BEP}.

Results of a heuristic method are provided in our paper.

Experiments

Methods:

Proposed framework:

Sigmoid: AUC maximization using **symmetric** sigmoid loss
Logistic: AUC maximization using **non-symmetric** logistic loss

$\ell_{\mathrm{sigmoid}}^{\mathrm{sym}}(z) = \frac{1}{1+\exp(z)}$
$\ell_{\log}(z) = \log(1+\exp(-z))$

Text feature baselines:

Maxent: maximum entropy classifier
NB: naïve bayes
PU-NB: variant of the NB that performs classification using positive and unlabeled data

GloVe baselines:

Randomforest: random forest on average word vectors
KNN: k-nearest neighbors on average word vectors

Zero-shot baselines:

GloveRanking: rank the score by average distance to relevant keywords
Voting: majority vote by keywords

Does threshold adjustment help?

F1-score: without adjustment

Methods	Subj	MPQA	AYI	20NG
Maxent	63.4 (0.31)	50.1 (0.22)	42.5 (0.35)	47.4 (0.05)
NB	73.7 (0.23)	53.8 (0.22)	**65.8** (0.42)	23.7 (0.25)
RandomForest	33.3 (0.00)	43.5 (0.20)	35.0 (0.20)	47.2 (0.00)
KNN	43.6 (0.23)	**51.0 (0.16)**	61.6 (0.43)	84.3 (0.26)

F1-score: with adjustment

Methods	Subj	MPQA	AYI	20NG
Maxent	**76.3 (0.24)**	**53.1 (0.20)**	**56.6 (0.36)**	**52.4 (0.25)**
NB	**76.3 (0.16)**	**54.3 (0.28)**	61.6 (0.38)	**58.4 (0.22)**
RandomForest	**75.1 (0.27)**	**62.4 (0.45)**	**64.5 (0.53)**	**89.6 (0.28)**
KNN	**63.6 (0.32)**	23.8 (0.00)	**65.5 (0.52)**	**86.7 (0.59)**

Threshold adjustment can improve the performance in most cases.

F1-score: threshold selection method with different $\hat{\pi}$:

$\pi_{\mathrm{Subj}} = 0.50$

Dataset	Methods	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Subj	Sigmoid	43.3 (0.17)	51.2 (0.24)	63.9 (0.34)	72.5 (0.27)	77.9 (0.37)	**80.1 (0.38)**	78.7 (0.32)	73.7 (0.29)
	Maxent	37.1 (0.15)	41.9 (0.18)	52.5 (0.27)	62.1 (0.19)	70.6 (0.27)	**76.3 (0.24)**	74.5 (0.25)	64.1 (0.25)
	RandomForest	42.1 (0.22)	49.9 (0.15)	61.6 (0.22)	69.1 (0.26)	73.5 (0.29)	**75.1 (0.27)**	73.5 (0.17)	69.0 (0.21)
	GloVe Ranking	43.1 (0.24)	50.5 (0.25)	61.9 (0.25)	69.3 (0.27)	73.6 (0.18)	**74.5 (0.13)**	72.8 (0.18)	68.0 (0.16)
20NG	Sigmoid	79.9 (0.23)	**91.2 (0.18)**	78.4 (0.20)	68.2 (0.15)	60.0 (0.14)	52.7 (0.13)	45.3 (0.15)	37.8 (0.13)
	Maxent	51.5 (0.19)	**52.3 (0.24)**	51.5 (0.21)	49.5 (0.16)	46.6 (0.15)	43.0 (0.16)	38.1 (0.14)	32.9 (0.15)
	RandomForest	79.4 (0.33)	**89.8 (0.29)**	78.5 (0.17)	68.2 (0.15)	59.8 (0.16)	52.6 (0.13)	45.4 (0.12)	37.7 (0.16)
	GloVe Ranking	79.2 (0.34)	**90.7 (0.17)**	78.2 (0.22)	67.9 (0.19)	59.9 (0.14)	52.3 (0.11)	45.0 (0.10)	37.6 (0.10)

$\pi_{\mathrm{20NG}} = 0.11$

The closer $\hat{\pi}$ to π, the better F1-score.

Four evaluation metrics with adjusted threshold:

(Prec@100 and AUC do not need threshold)

Dataset	Evaluation	Proposed framework		Text-feature baselines			GloVe-feature baselines		Zero-shot baselines		Oracle	
		Sigmoid	Logistic	PU-NB	NB	Maxent	RandomForest	KNN	GloVeRanking	Voting	O-Maxent	O-Sigmoid
Subj	AUC	**88.1 (0.35)**	84.1 (0.30)	55.4 (0.13)	85.0 (0.18)	84.6 (0.20)	82.4 (0.27)	73.6 (0.29)	81.7 (0.19)	70.2 (0.24)	97.4 (0.06)	93.6 (0.11)
	F_1	**80.1 (0.38)**	76.0 (0.32)	47.1 (0.20)	76.3 (0.16)	76.3 (0.24)	75.1 (0.27)	63.6 (0.32)	74.5 (0.13)	63.5 (0.18)	92.0 (0.13)	86.4 (0.14)
	ACC	**80.1 (0.38)**	76.0 (0.32)	55.0 (0.13)	76.3 (0.16)	76.3 (0.24)	75.1 (0.27)	65.0 (0.28)	74.5 (0.13)	64.1 (0.18)	92.0 (0.13)	86.4 (0.14)
	Prec@100	**96.3 (0.60)**	**95.1 (0.60)**	0.9 (0.09)	**95.9 (0.33)**	94.7 (0.39)	93.2 (0.50)	91.5 (0.59)	**95.2 (0.54)**	85.8 (0.93)	99.3 (0.15)	97.8 (0.27)
MPQA	AUC	**80.4 (0.44)**	78.7 (0.37)	52.1 (0.27)	56.4 (0.31)	56.7 (0.23)	69.1 (0.55)	60.1 (0.23)	63.6 (0.26)	56.0 (0.12)	78.3 (0.25)	86.8 (0.18)
	F_1	**71.7 (0.44)**	69.8 (0.31)	46.7 (0.23)	54.3 (0.28)	53.1 (0.20)	62.4 (0.45)	23.8 (0.00)	57.5 (0.17)	23.8 (0.00)	69.8 (0.19)	77.9 (0.22)
	ACC	**75.6 (0.39)**	74.0 (0.27)	47.1 (0.24)	62.4 (0.28)	58.4 (0.17)	67.4 (0.39)	31.2 (0.00)	63.3 (0.17)	31.2 (0.00)	72.8 (0.20)	81.0 (0.19)
	Prec@100	**81.5 (0.97)**	77.5 (1.02)	10.8 (3.37)	69.5 (0.86)	63.8 (1.93)	76.9 (1.06)	78.7 (0.80)	50.6 (0.60)	74.7 (0.69)	94.8 (0.46)	90.5 (0.52)
AYI	AUC	**76.0 (0.41)**	**75.6 (0.43)**	60.5 (0.39)	71.2 (0.41)	60.7 (0.46)	70.1 (0.55)	72.5 (0.39)	62.4 (0.53)	61.0 (0.33)	84.6 (0.32)	81.1 (0.40)
	F_1	**69.3 (0.36)**	**68.8 (0.40)**	58.9 (0.47)	61.6 (0.38)	56.6 (0.36)	64.5 (0.53)	65.5 (0.52)	58.7 (0.51)	33.5 (0.00)	76.8 (0.37)	73.0 (0.39)
	ACC	**69.3 (0.36)**	**68.8 (0.40)**	60.1 (0.41)	62.5 (0.35)	56.8 (0.36)	64.6 (0.53)	65.8 (0.44)	58.7 (0.51)	50.5 (0.00)	76.9 (0.37)	73.0 (0.39)
	Prec@100	**87.5 (0.55)**	**87.5 (0.62)**	74.5 (2.20)	85.1 (0.71)	70.2 (1.00)	77.2 (0.99)	82.5 (0.69)	72.4 (0.91)	79.2 (0.87)	95.6 (0.47)	90.1 (0.73)
20NG	AUC	96.4 (0.12)	96.0 (0.15)	N/A	77.1 (0.21)	57.6 (0.32)	**96.8 (0.16)**	94.7 (0.16)	95.0 (0.17)	62.9 (0.22)	65.5 (0.46)	99.0 (0.05)
	F_1	**90.8 (0.20)**	**90.6 (0.21)**	N/A	58.4 (0.22)	52.4 (0.25)	89.6 (0.28)	86.7 (0.59)	**90.5 (0.18)**	9.6 (0.00)	56.8 (0.29)	94.1 (0.14)
	ACC	**96.5 (0.08)**	**96.4 (0.08)**	N/A	70.2 (0.31)	81.6 (0.11)	96.1 (0.10)	94.5 (0.35)	**96.4 (0.07)**	10.6 (0.00)	83.5 (0.11)	97.8 (0.05)
	Prec@100	**99.5 (0.15)**	**99.1 (0.24)**	N/A	0.4 (0.11)	17.6 (0.77)	**99.5 (0.15)**	97.6 (0.38)	97.5 (0.28)	85.2 (1.03)	32.0 (1.31)	99.9 (0.07)

Fully-labeled data are given

References

[1] Liu, B., Li, X., Lee, W. S., and Yu, P. S. Text classification by labeling words. In AAAI, 2004.
[2] Druck, G., Mann, G., and McCallum, A. Learning from labeled features using generalized expectation criteria. SIGIR, 2008.
[3] Settle, B. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. EMNLP, 2011.
[4] Jin, Y., Wanvarie, D., and Le, P. Combining lightly-supervised text classification models for accurate contextual advertising. IJCNLP, 2017..
[5] Charoenphakdee, N., Lee, J., and Sugiyama, M. On symmetric losses for learning from corrupted labels. ICML, 2019.
[6] Kato, K., Teshima, T., and Honda, J. Learning from positive and unlabeled data with a selection bias. ICLR, 2019.