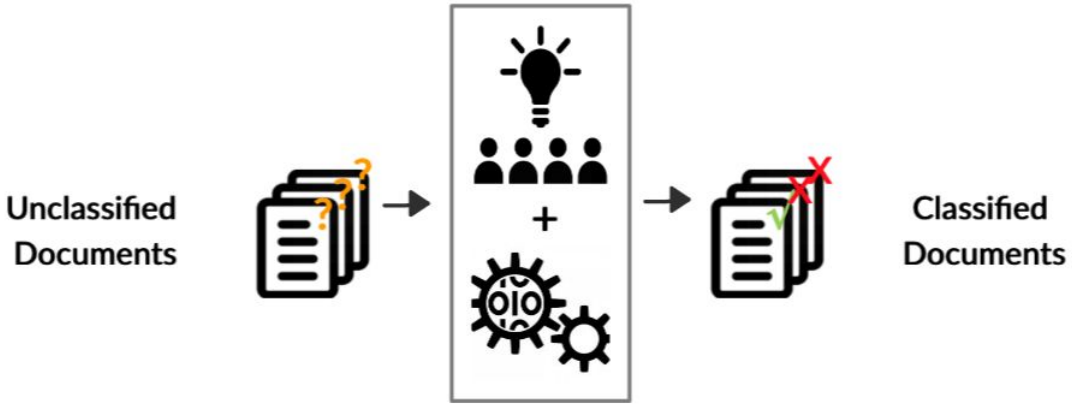
Towards Reliable Hybrid Human-Machine Classifiers

Burcu Sayin Günel

Hybrid Classification

- Classification problems solved via the combined contribution of **humans** and **machines**.
- **Train machines efficiently vs use humans efficiently**



Nearly every end to end ML-powered application is “hybrid”

Machines are not always reliable

Feds open formal investigation following Tesla 'Autopilot' crashes

By Will Feuer

August 16, 2021 | 9:50am | Updated

Since 2018, there have been 11 crashes in which Teslas on Autopilot or Traffic Aware Cruise Control have hit emergency vehicles.

Metrics

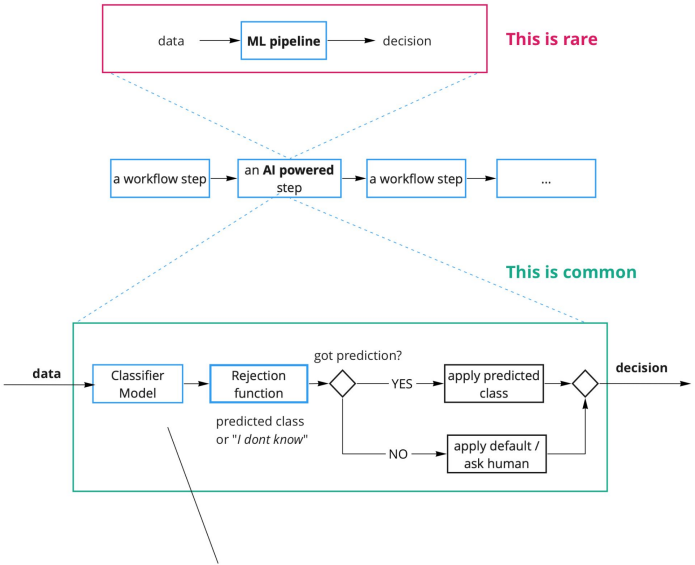
Inference, confidence, accuracy

Trained
ML model

75% cat 20% bat 5% dog	Inference	Confidence	Accuracy
	CAT	75%	1 (Assuming we have only one inference.)

In case the confidence threshold is 80%; this inference is not trustable.

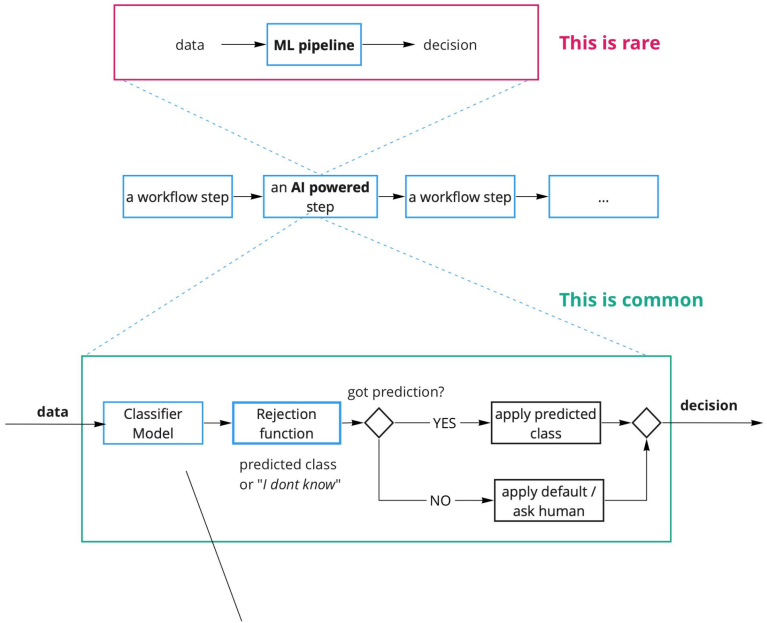
Typical implementation of ML models into an enterprise workflow.



1. A typical implementation of the rejection function is to filter based on a confidence threshold
2. The above assumes that the classifier is trained independently of the rejection logic. This does not have to be - the classifier can be aware of the cost, in which case it becomes less "general"

REF1: Burcu Sayin, Jie Yang, Andrea Passerini, and Fabio Casati. The science of rejection: A research area for human computation. In The 9th AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2021. AAAI Press, 2021.

Observations



1. A typical implementation of the rejection function is to filter based on a confidence threshold
2. The above assumes that the classifier is trained independently of the rejection logic. This does not have to be - the classifier can be aware of the cost, in which case it becomes less "general"

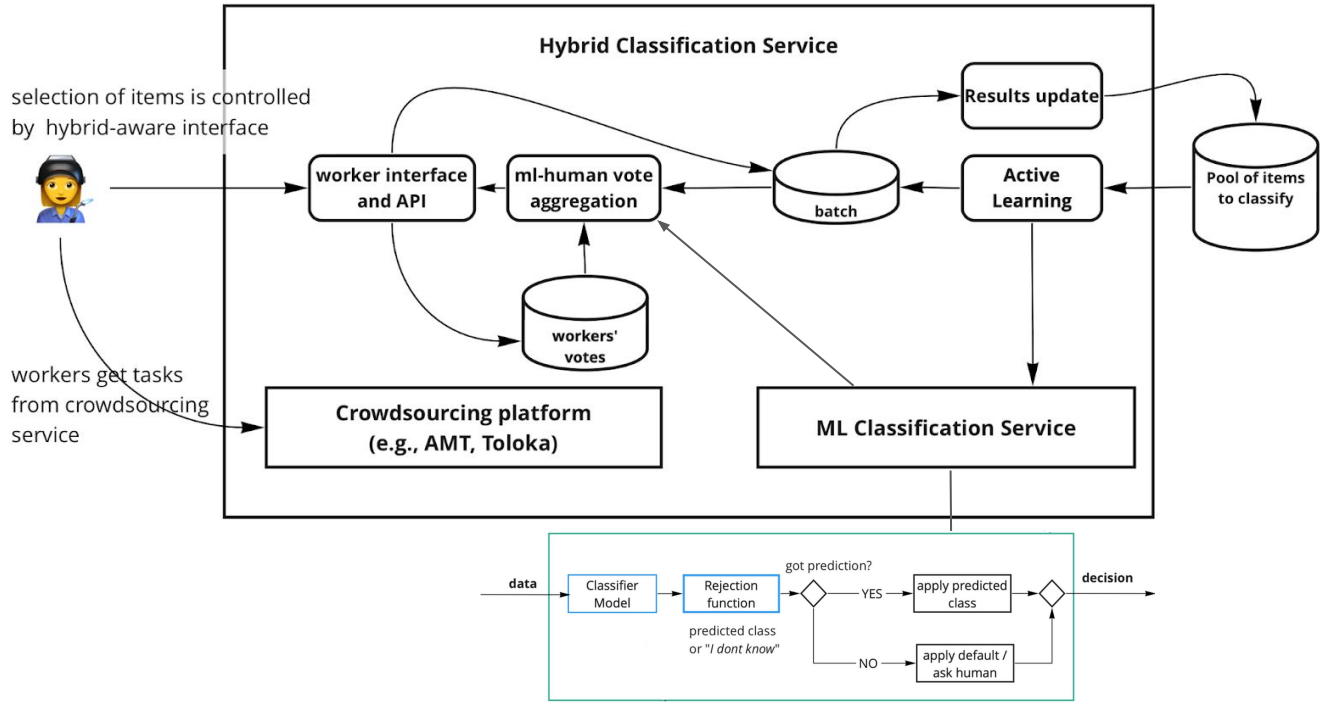
1. The threshold and the system behavior depend on **the ``cost" of machine errors** and its relation to **the cost of a rejection** and **the value of a correct machine prediction**.
2. Use case drives the value of ML models
3. If we have a well-calibrated model with arbitrarily bad accuracy, we can still get value from it. The better we are able to identify subset of items for which our model is calibrated, the lower is the cost for our deployment in an AI workflow.
4. We really care about the fact that the machine should be aware of what it knows and give alert when it does not know.


"The only true wisdom is in knowing you know **nothing**." – *Socrates*
Can a machine tell when it does not know?
Knowing if we can **trust** a hybrid human-machine classification service **is the key**.

Problem

How to design and implement reliable hybrid human-machine classification services?

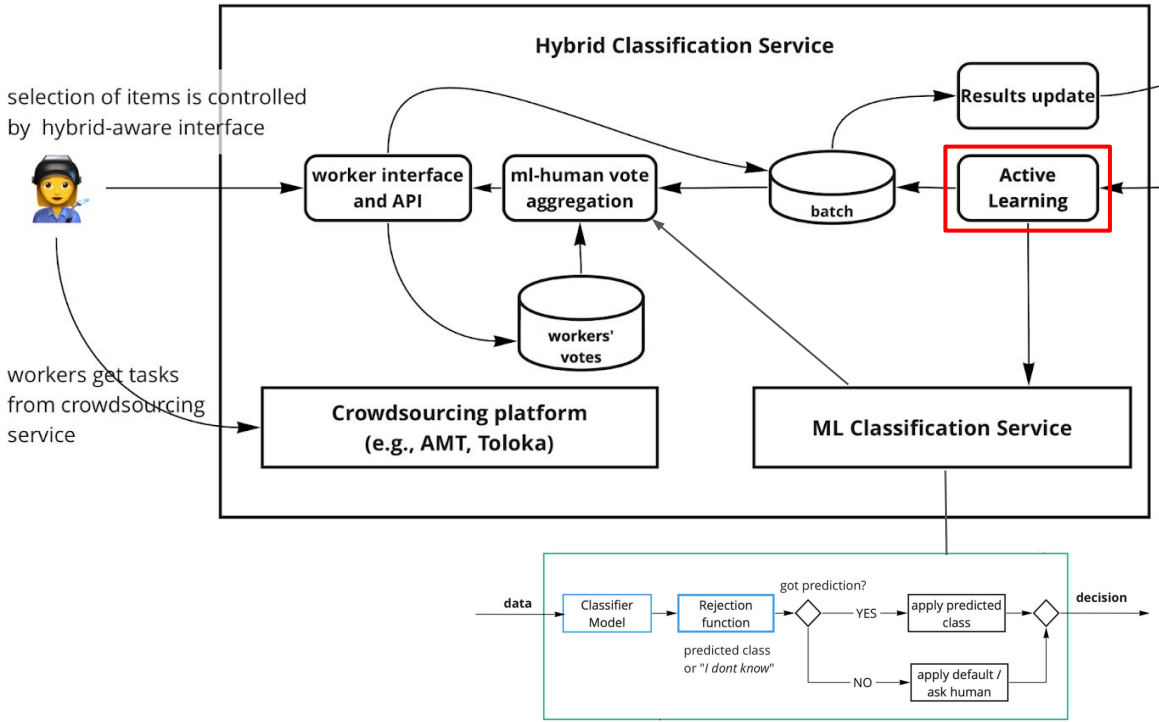
Approach and Architecture



REF2: B. Sayin et al., "Crowd-Powered Hybrid Classification Services: Calibration is all you need," 2021 IEEE International Conference on Web Services (ICWS), 2021, pp. 42-50.

REF3: E. Krivosheev et al. Active hybrid classification. In Proceedings of NeurIPS Workshop on Human in the Loop Dialogue Systems, 2020.

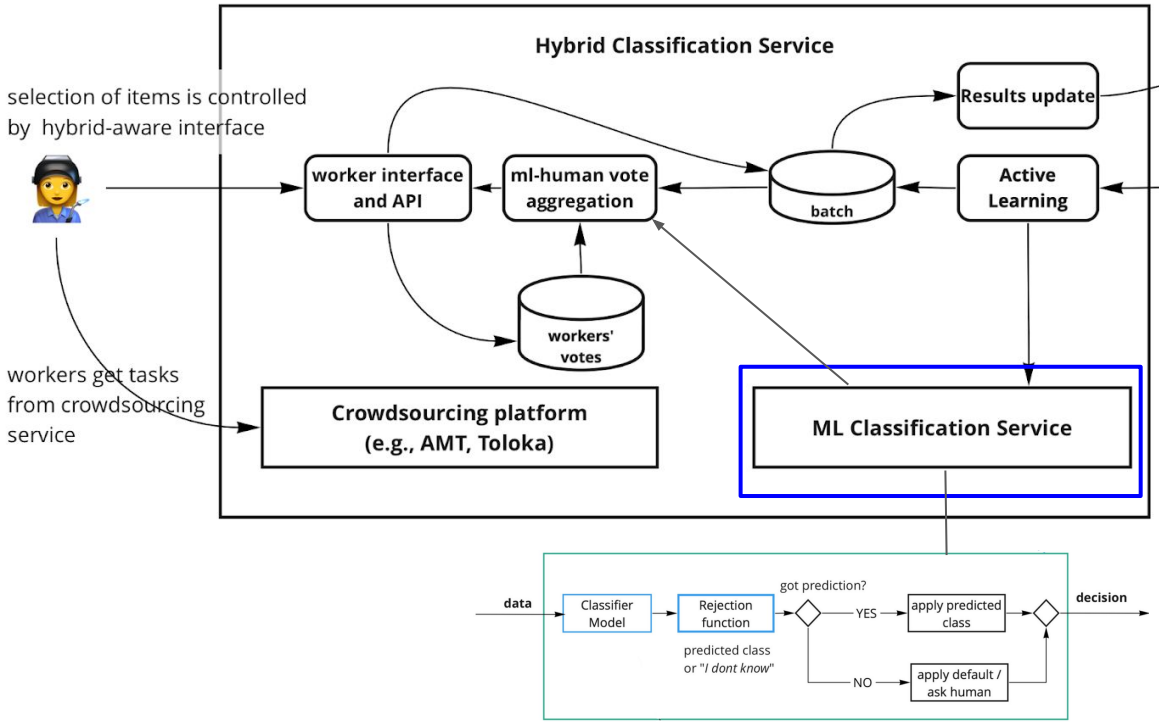
Research Questions



RQ1. How to use active learning (AL) in hybrid classification contexts? Are the existing AL strategies cost-effective or do we need novel cost-aware AL methods for hybrid classification services?

Burcu Sayin, Evgeny Krivosheev, Jie Yang, Andrea Passerini, and Fabio Casati. **A review and experimental analysis of active learning over crowdsourced data**. In *Journal of Artificial Intelligence Review*, Volume 54, pp. 5283-5305, 2021.

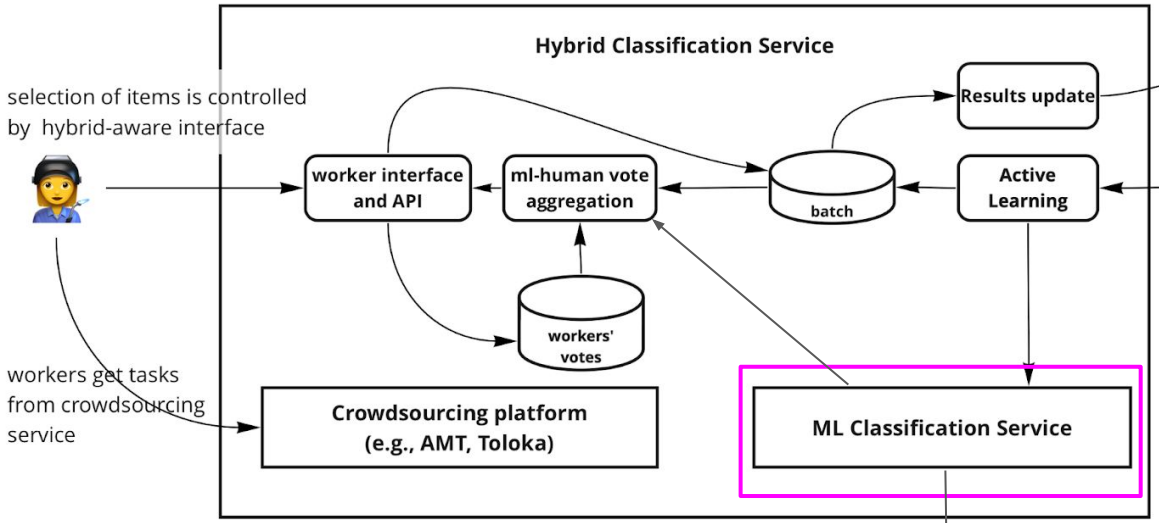
Research Questions

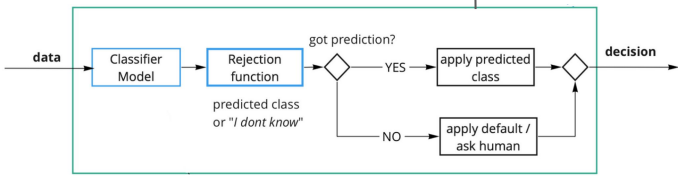


RQ2. How does the model calibration affect the performance of ML models in hybrid classification contexts? And, how can we obtain well-calibrated hybrid classifiers?

Burcu Sayin, Evgeny Krivosheev, Jorge Ramírez, Fabio Casati, Ekaterina Taran, Veronika Malanina, and Jie Yang. **Crowd-Powered Hybrid Classification Services: Calibration is all you need**. In Proceedings of the IEEE International Conference on Web Services (ICWS), pp. 42-50. 2021.

Research Questions





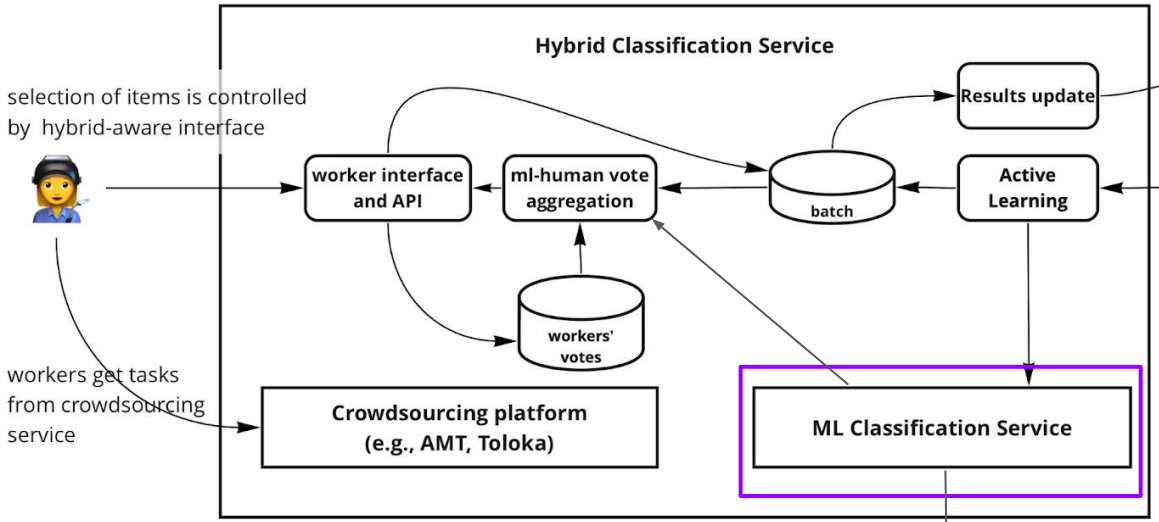
RQ3. How to effectively characterize machine learning failures?

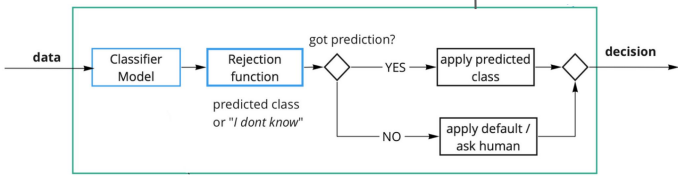
Shahin Sharifi, Sihang Qiu, **Burcu Sayin**, Agathe Balayn, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. (2022) **Perspective: Leveraging Human Understanding for Identifying and Characterizing Image Atypicality**. (Under Review)

As a part of Google's CATS4ML 2021 Challenge:

cats4ml.humancomputation.com/

Research Questions

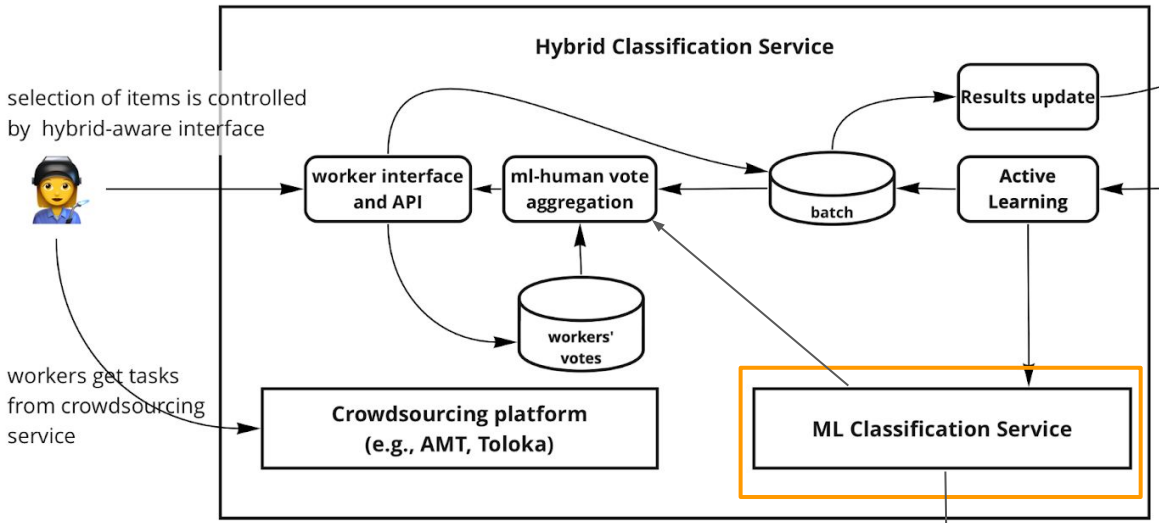




RQ4. What are the proper metrics for the cost of using ML with rejection?

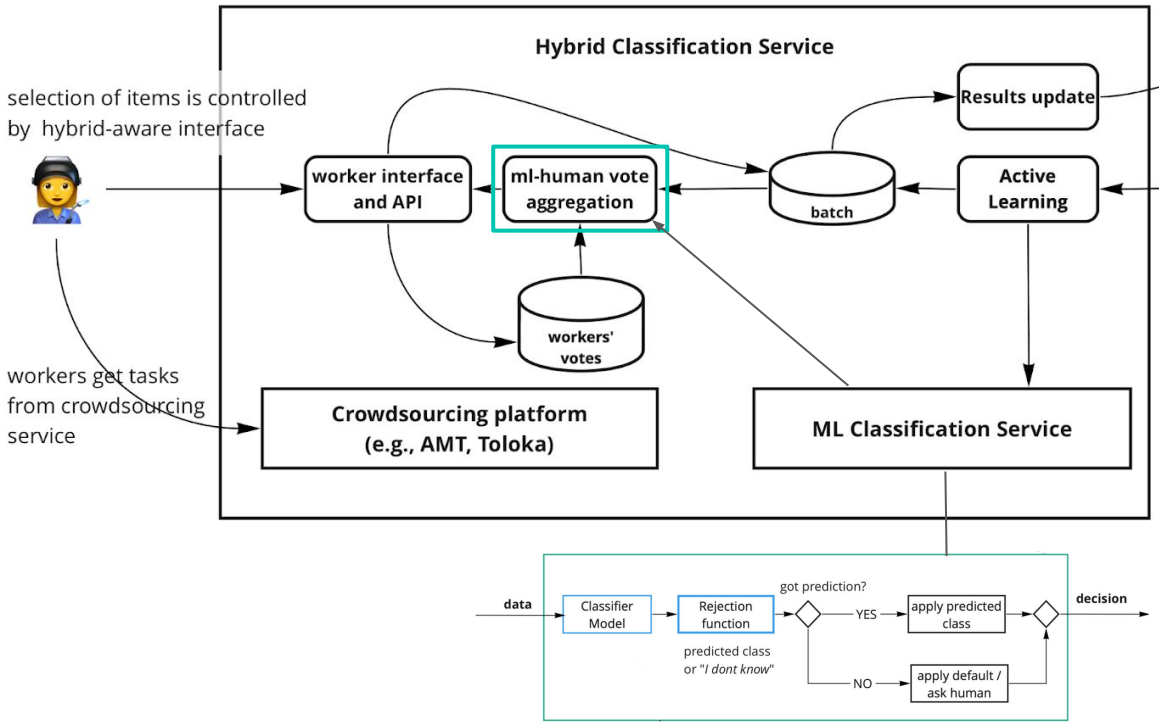
Burcu Sayin, Jie Yang, Andrea Passerini, and Fabio Casati. **The science of rejection: A research area for human computation.** In The 9th AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2021. AAAI Press, 2021.

Research Questions



RQ5. How to build the rejector?

Research Questions



RQ6. How to efficiently combine crowdsourcing and machine intelligence?

Evgeny Krivosheev, **Burcu Sayin**, Alessandro Bozzon, and Zoltan Szlavik. **Active learning from crowd in document screening**. In Crowd Science Workshop at 34th Conference on Neural Information Processing Systems (*Neur*IPS 2020).

Research Questions - Summary

RQ1. How to use **active learning** (AL) in hybrid classification contexts?

Are the existing AL strategies cost-effective or do we need novel cost-aware AL methods for hybrid classification services?

RQ2. How does the model **calibration** affect the performance of ML models in hybrid classification contexts?

And, how can we obtain well-calibrated hybrid classifiers?

RQ3. How to effectively characterize machine learning **failures**?

RQ4. What are the proper **metrics** for the cost of using ML with rejection?

RQ5. How to build the **rejector**?

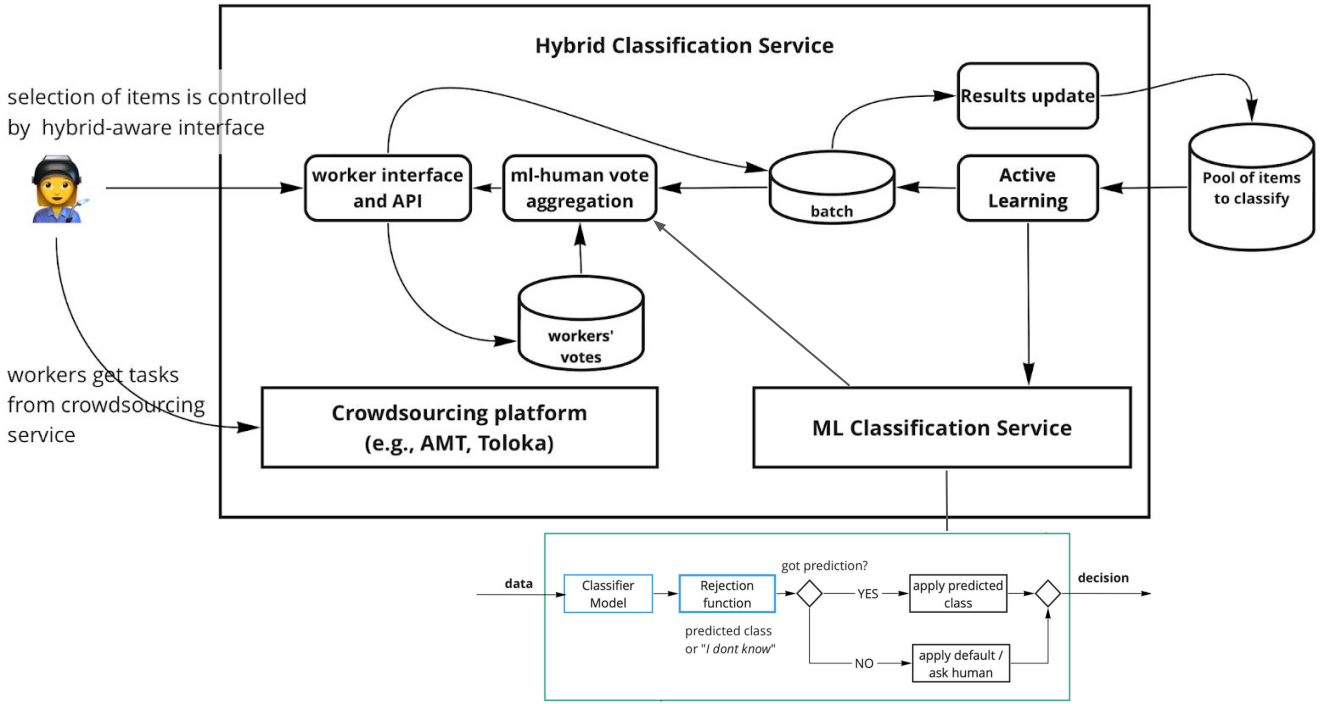
RQ6. How to efficiently combine crowdsourcing and machine intelligence?

RQ 2

The centrality of calibration in hybrid classification

Burcu Sayin, Evgeny Krivosheev, Jorge Ramírez, Fabio Casati, Ekaterina Taran, Veronika Malanina, and Jie Yang. **Crowd-Powered Hybrid Classification Services: Calibration is all you need**. In Proceedings of the IEEE International Conference on Web Services (ICWS), pp. 42-50. 2021.

Hybrid Classification Service Architecture



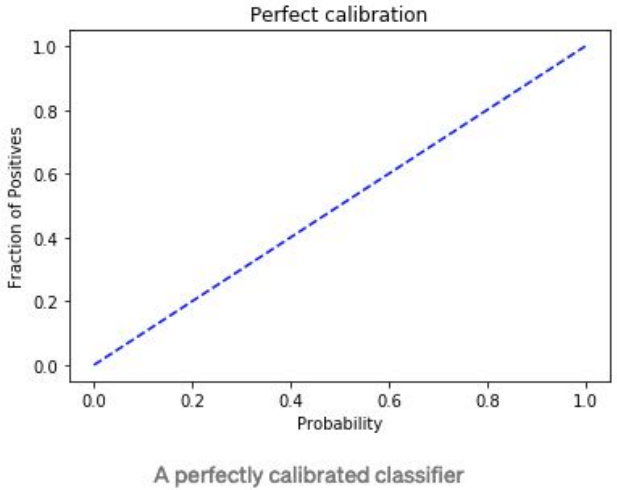
REF2: B. Sayin et al., "Crowd-Powered Hybrid Classification Services: Calibration is all you need," 2021 IEEE International Conference on Web Services (ICWS), 2021, pp. 42-50.

REF3: E. Krivosheev et al. Active hybrid classification. In Proceedings of NeurIPS Workshop on Human in the Loop Dialogue Systems, 2020.

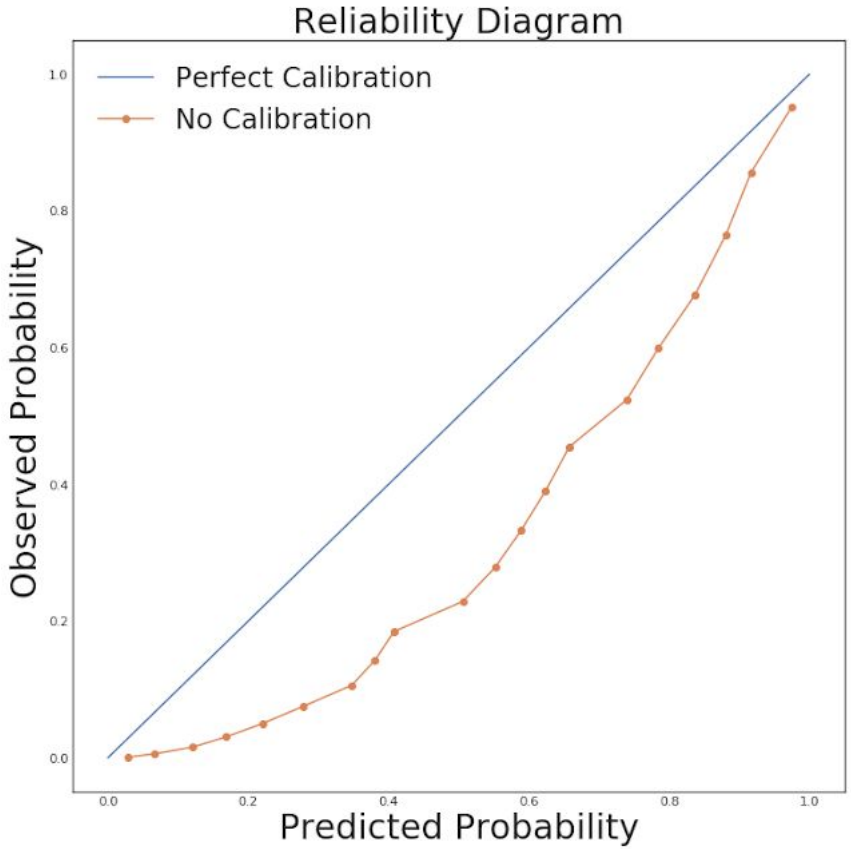
Calibration

Data point	Label
x_1	1
x_2	1
x_3	1
x_4	1
x_5	0
x_6	1
x_7	0
x_8	1
x_9	1
x_{10}	0

The fraction of positive points is 0.7 (**observed probability**)
What about the **predicted probability?**



medium.com/analytics-vidhya/calibration-in-machine-learning-e7972ac93555



www.unofficialgoogledatascience.com/2021/04/why-model-calibration-matters-and-how.html

Measuring Expected Calibration Error (ECE) for a Multi-Class Classification Model

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{n_b}{N} \left|\mathrm{acc}(b) - \mathrm{conf}(b)\right|$$

item	output pps					largest pp	predicted	target	result	
[0]	0.21	0.58	0.09	0.12		0.58	1	1	correct	
[1]	0.32	0.26	0.05	0.37		0.37	3	0	wrong	
[2]	0.28	0.25	0.26	0.21		0.28	0	0	correct	
[3]	0.02	0.93	0.03	0.02		0.93	1	1	correct	
[4]	0.22	0.27	0.20	0.31		0.31	3	1	wrong	
[5]	0.19	0.15	0.61	0.05		0.61	2	2	correct	
[6]	0.84	0.05	0.03	0.08		0.84	0	3	wrong	
[7]	0.18	0.09	0.70	0.03		0.70	2	2	correct	
[8]	0.11	0.44	0.32	0.13		0.44	1	1	correct	
[9]	0.20	0.26	0.15	0.39		0.39	3	3	correct	
	0	1	2	3						
				bin		count	accuracy	avg pp	|acc - a pp|	count *
				1	0.0 to 0.33	2	1/2 = 0.500	0.295	0.205	0.410
				2	0.34 to 0.66	5	4/5 = 0.800	0.462	0.338	1.690
				3	0.67 to 1.0	3	2/3 = 0.667	0.823	0.156	0.468
						10				2.568
									ECE =	0.257

Figure is taken from pureai.com/articles/2021/03/03/ml-calibration.aspx

Related works

- Szegedy et. al. (2016) propose to adopt **label smoothing**, where "hard" (1-0) class labels used in cross-entropy loss are smoothed into a probability distribution across classes
    - Temperature scaling (Guo et al. 2017): multiplying the logits by a scalar before the softmax operator.
    - probability amortization in output targets can bring extra noises and prior art does not provide insights in how to set label smoothing hyper-parameters.
- An alternative approach; creating an empirical distribution over crowd votes -**soft target approach**- (Wulczyn et al, 2017; Aung and Whitehill, 2018). -> **None has investigated the effect of label fusion methods on model training.**

- We study **the effect of label smoothing and soft targets on model calibration in NLP tasks when training labels are crowdsourced**.
- **We propose alternative soft-target approaches to improve model calibration.**

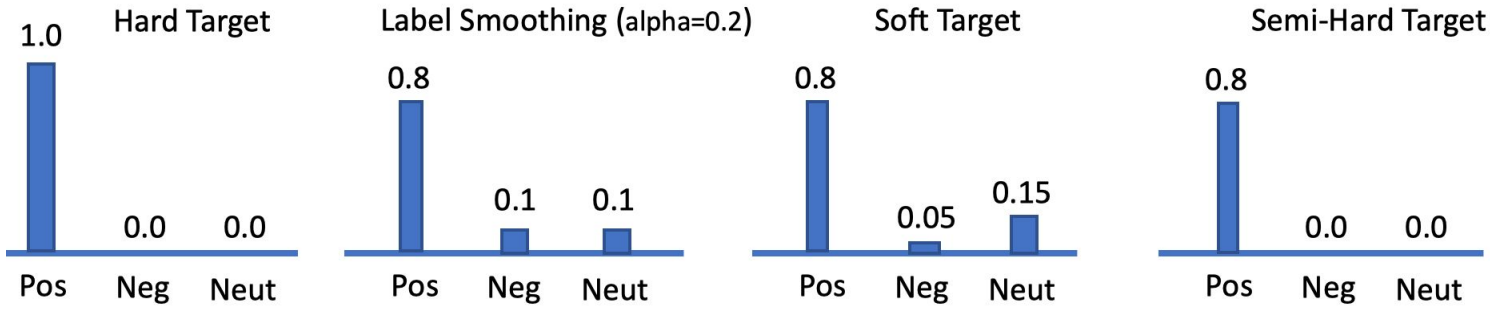
Proposed soft target approaches

1. **Soft targets (Soft): F** is a function that takes crowd labels as input, and outputs a set of workers' accuracies **A** and a probability assignment $\boldsymbol{\pi}_f$ (via label fusion techniques) over the classes for every sample:

$$\mathcal{F}: \ crowd\ labels \rightarrow \langle \pi_f, A \rangle$$

2. **Soft-hard targets (sHard):** one-side smoothing that of the most likely label as identified by the fusion method

$$\pi_{sh}(l|x) = \begin{cases} \pi_f(l|votes, F), & l = l^* \\ 0, & l \neq l^* \end{cases}$$



Different target mass functions for classification problem with three classes (Pos- positive, Neg- Negative, Net- Neutral classes). The Hard Target is equivalent to one-hot label encoding, the distribution for Soft Target is obtained from crowd votes via a label fusion method (e.g., MV or DS).

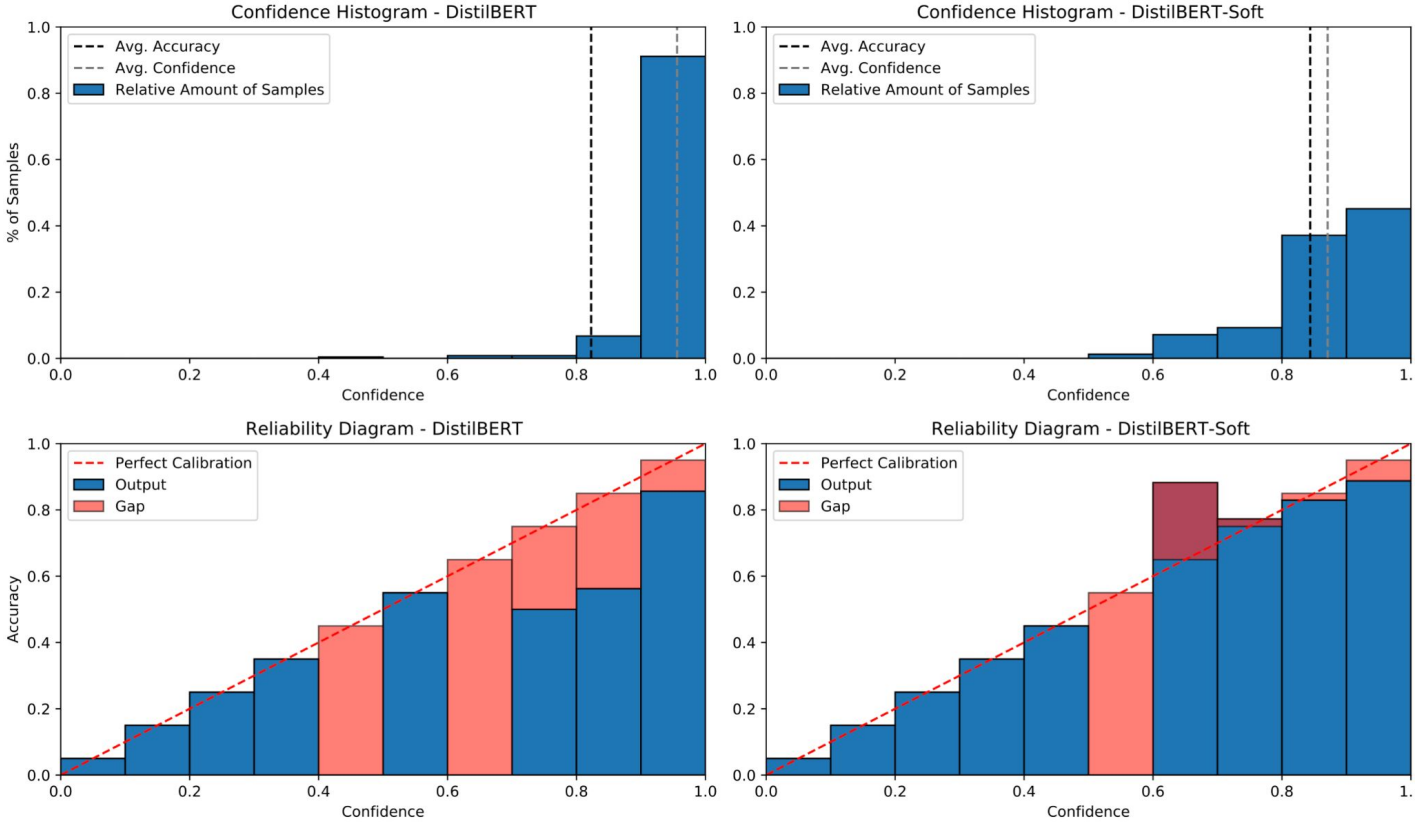
Experimental Setup

- 5 binary, 8 multi-class datasets (3 of them have individual votes)
- We evaluated:
  - a simple one-layer neural network (NN1) with text vectorized via tf-idf
  - fine-tuned DistilBERT model (D-BERT) with 6 layers, 768 hidden dim, 12 heads, and 65M parameters
- We trained them with the *cross-entropy loss* using:
  - hard targets (one-hot encoded labels)
  - label smoothing
  - the proposed soft targets
- We tested the impact of three label fusion methods:
  - Majority Voting (MV)
  - Dawid-Skene (D&S), which models worker reliability
  - *GLAD*, which further considers the task difficulty

Results - Performance of NN1 and D-BERT with targets obtained from different fusion methods

Model	Movie-Reviews		Reuters-21578		Amazon-Reviews	
	ECE	F1	ECE	F1	ECE	F1
NN1-Hard labels (MV)	4,2	58,5	**4,7**	63,6	**11,0**	**95,8**
NN1-Soft (MV)	**3,1**	**59,8**	5,3	65,1	12,6	**95,8**
NN1 (α=0.05, MV)	8,2	53,5	6,1	64,6	17,3	95,7
NN1 (α=0.1, MV)	10,4	55,3	5,4	**65,5**	20,8	**95,8**
NN1-Hard labels (DS)	14,3	57,6	3,6	**72,4**	11,0	**95,8**
NN1-Soft (DS)	**8,7**	**58,5**	4,5	70,9	**10,6**	95,7
NN1 (α=0.05, DS)	9,5	56,1	**2,9**	70,8	17,3	**95,8**
NN1 (α=0.1, DS)	**8,7**	55,2	4,7	69,7	20,8	**95,8**
NN1-Hard labels (GLAD)	**4,5**	**57,7**	**3,1**	67,1	**10,5**	**96,0**
NN1-Soft (GLAD)	6,6	56,9	4,1	**70,6**	10,9	95,7
NN1 (α=0.05, GLAD)	9,3	52,5	3,7	68,6	17,1	95,8
NN1 (α=0.1, GLAD)	7,0	55,7	4,4	69,6	20,6	95,8
D-BERT-Hard labels (MV)	36,5	56,3	-	-	5,9	**93,0**
D-BERT-Soft (MV)	**21,1**	54,9	-	-	**5,3**	92,4
D-BERT (α=0.05, MV)	25,7	**57**	-	-	7,3	92,4
D-BERT (α=0.1, MV)	21,6	56,7	-	-	11,5	**93,0**
D-BERT-Hard labels (DS)	34,2	56,5	-	-	3,3	92,7
D-BERT-Soft (DS)	34,1	53,6	-	-	**2,0**	**93,1**
D-BERT (α=0.05, DS)	**23,0**	**58,5**	-	-	7,7	92,8
D-BERT (α=0.1, DS)	31,6	49,4	-	-	10,3	**93,1**
D-BERT-Hard labels (GLAD)	36,1	56,6	-	-	9,2	91,1
D-BERT-Soft (GLAD)	**20,4**	**58,2**	-	-	**2,1**	**94,1**
D-BERT (α=0.05, GLAD)	29,5	54,9	-	-	7,7	92,8
D-BERT (α=0.1, GLAD)	24,5	55,6	-	-	12,8	93,4

The proposed soft target method improved ECE (up to 15.7%) across all datasets.



The effect of soft targets on probability calibration for DistilBERT on Death-in-India dataset

Results - Performance of NN1 and D-BERT with sHard targets obtained from different fusion methods.

Model	Movie-Reviews		Reuters-21578		Amazon-Reviews	
	ECE	F1	ECE	F1	ECE	F1
NN1-sHard (MV)	4,3	**60,2**	5,1	**64,3**	**10,3**	95,5
NN1-sHard (DS)	**10,5**	**58,2**	3,7	72,3	**10,4**	95,7
NN1-sHard (GLAD)	5,7	**58,6**	3,4	67,1	10,8	95,8
D-BERT-sHard (MV)	**35,9**	**57,4**	-	-	**4,9**	92,5
D-BERT-sHard (DS)	**30,0**	**57,9**	-	-	3,3	92,1
D-BERT-sHard (GLAD)	**35,2**	56,4	-	-	**5,1**	**93,3**

sHard improves ECE on 2 datasets that have individual crowd votes while remaining comparable on 1 dataset

Results - Average improvement of soft targets compared to hard targets (in %)

Model	F1	ECE	Model	F1	ECE
NN1-sHard	0.04	**-1.7**	D-BERT-sHard	**0.6**	-1.8
NN1-Soft	0.1	-0.7	D-BERT-Soft	0.3	**-7.2**
NN1 (α=0.05)	**0.3**	0.9	D-BERT (α=0.05)	-3.6	-1.1
NN1 (α=0.1)	-0.4	1.5	D-BERT (α=0.1)	-0.3	-5.2

Our proposed Soft and sHard target methods substantially improves the model calibration.

Take-away messages

1. Calibration is central in contexts where the loss function is skewed and the cost of errors is high compared to the cost of asking humans.
2. Soft and soft-hard targets help improving the calibration of text classification with crowdsourced data
3. The effect on calibration error and the benefits of the proposed approach are manifest for deep models, that are known to be more affected by calibration issues

References

1. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. http://arxiv.org/abs/1512.00567
2. Chuan Guo,Geoff Pleiss,Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In Proceedings of the 34th International Conference on Machine Learning - Volume 70 (Sydney, NSW, Australia) (ICML'17). JMLR.org, 1321–1330.
3. Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal attacks seen at scale. In Proceedings of the 26th International Conference on World Wide Web, pp. 1391–1399.
4. Arkar Min Aung and Jacob Whitehill. 2018. Harnessing Label Uncertainty to Improve Modeling: An Application to Student Engagement Recognition. In FG. 166–170.

Thank you for your attention

Contact: burcusayinn@gmail.com
burcu.sayin@unitn.it