Average Jane, Where Art Thou? – Recent Avenues in Efficient Machine Learning Under Subjectivity Uncertainty

Georgios Rizos[1] and Björn W. Schuller[1,2,3(✉)]

[1] GLAM, Imperial College London, London SW7 2AZ, UK
{rizos,schuller}@ieee.org
[2] Chair of EIHW, University of Augsburg, 86159 Augsburg, Germany
[3] audEERING, 82205 Gilching, Germany

Abstract. In machine learning tasks an actual 'ground truth' may not be available. Then, machines often have to rely on human labelling of data. This becomes challenging the more subjective the learning task is, as human agreement can be low. To cope with the resulting high uncertainty, one could train individual models reflecting a single human's opinion. However, this is not viable, if one aims at mirroring the general opinion of a hypothetical 'completely average person' – the 'average Jane'. Here, I summarise approaches to optimally learn efficiently in such a case. First, different strategies of reaching a single learning target from several labellers will be discussed. This includes varying labeller trustability and the case of time-continuous labels with potential dynamics. As human labelling is a labour-intensive endeavour, active and cooperative learning strategies can help reduce the number of labels needed. Next, sample informativeness can be exploited in teacher-based algorithms to additionally weigh data by certainty. In addition, multi-target learning of different labeller tracks in parallel and/or of the uncertainty can help improve the model robustness and provide an additional uncertainty measure. Cross-modal strategies to reduce uncertainty offer another view. From these and further recent strategies, I distil a number of future avenues to handle subjective uncertainty in machine learning. These comprise bigger, yet weakly labelled data processing basing amongst other on reinforcement learning, lifelong learning, and self-learning. Illustrative examples stem from the fields of Affective Computing and Digital Health – both notoriously marked by subjectivity uncertainty.

Keywords: Machine learning · Uncertainty · Subjectivity · Active learning · Cooperative learning

1 Subjectivity and AI

In many machine learning applications of interest, the ground truth reflects some inherently human-centric capacity, like affect [1], or corresponds to an

expert assessment, as in the case of health informatics [2]. In such cases, ground truth cannot be collected automatically via crawling or sensors, and **a human annotation step must be injected in the machine learning pipeline**.

Considering, for example, the differing reports made by experienced healthcare professionals when assessing a medical image [3], it is not hard to imagine that certain annotation tasks can exhibit great subjectivity. Given a photo of a man wearing a slight frown, is he sad, angry, or is this maybe his natural expression? What about a photo of a woman [4]? Is the gender of the person making the assessment [5] relevant? What about their age or cultural background [6]? In the online social media setting, is the mention of curse words (or even slurs) in a tweet considered offensive if it is used between friends [7]? Or, in the age of alternative facts, how can one be certain that an online post does not contain fake news [8]? Perhaps the task under consideration is **inherently ambiguous**, as in asking whether a scene depicted in a photo is warm or cold [9], or **simply difficult even for experts**, as in the identification of volcanoes on a photo of a $75 \cdot 75\,\mathrm{km}^2$ surface patch of Venus [10]. Such ambiguities become exacerbated when the level of expertise and trustworthiness of the annotators enter consideration [11].

In subjective perception studies like those in the above non-comprehensive list, we observe that if we were to ask multiple humans to annotate a single sample, we would often receive disagreeing evaluations. In fact, we typically would *require* multiple raters in an attempt to eliminate the possible bias of one single annotator. We are now faced with a different problem however: **our ground truth, instead of providing clear answers, introduces uncertainty, and disagreement as well; which opinion is correct, if any?** Here we present an overview of approaches utilised to address the issue of ground-truth uncertainty due to subjectivity, a discussion of current state-of-the-art methods, persistent challenges, and an outline of promising future directions.

2 The Dimensions of AI for Subjective Data

The presence of multiple human evaluations per sample forms a constellation of challenges and opportunities, and certain approaches to address some of the former may fail to exploit the latter. Each approach is delineated by the decisions made to accommodate a series of problem dimensions. We discuss the principal dimension that allows for a more high-level clustering in the following subsection.

2.1 Adapting the Labels vs. Adapting the Algorithms

We fundamentally must make a decision on whether we desire to work on the traditional setting, in which samples assume **hard labels** after a fusion of the original, distinct opinions, or we want to introduce the additional modelling complexity to accommodate the particularities of handling subjective data.

We treat the two extreme philosophies as **treating subjectivity as a problem** (see Sect. 3), or **embracing subjectivity and potentially leveraging the opportunities it offers** for better learning (see Sect. 4), respectively.

2.2 The Many Considerations of Working with Subjective Data

There are many permutations in the placement of relevant approaches with respect to the below dimensions, and in certain cases a middle-way is proposed. **It is important to keep in mind what assumptions are implicitly being made by the methods summarised in this overview study**.

Subjectivity as a Source of Information: Whereas rater disagreement does indeed inject noise in the ground-truth that requires extra steps to accommodate, it is also true that it can be quantified and also be used as an additional source of information or algorithm feature if quantified [12–14].

Bad Data or *Interesting* Data? Taking an example from affective computing, certain speech utterances may be characterised as being **prototypical**, i.e., they are clear examples of one emotion, or can be **ambiguous**, [15,16], i.e., at the fuzzy border of more than one emotions. If we subscribe to the first extreme, then we assume that a hard label is the true label of a sample, regardless of whether we use one, and that disagreeing labels constitute observation noise due to rater bias, that needs to be removed by means of denoising. Very often, **data that exhibit high rater disagreement are assumed to be less informative**, due to the inability of raters to come to a consensus. On the other hand, **disagreeing labels may actually capture a separate mode of a true soft label distribution**.

Feasibility of Modelling Raters: Further to the above, in certain cases we have knowledge of the set of labels produced by each (anonymised) rater, and we can use this information to estimate the trustworthiness of each either by proxy of inter-rater reliability, or by using an ensemble of models, where each corresponds to a different rater. Inversely, usually in the case of crowdsourcing, we either do not know which rater provided which label, or we simply have very little overlap between raters to compare their performances. In such cases, we can only attempt to model rater types [17,18].

Predicting With Uncertainty: Depending on the application, it may be desirable to provide a measure of subjectivity uncertainty (or, inversely, confidence) alongside the prediction of each test sample. Tasks that require risk-aware AI, may be related to healthcare, self-driving car technology, or simply cases where catastrophic performance translates to great financial costs. Being able to predict subjectivity uncertainty may also be the primary task of interest [3].

Interactive Learning: One approach to reducing the rater disagreement for a label, is to repeatedly label it, leading to the need for a thorough study of how the disagreement is treated and utilised in such a process. Furthermore, even in regular **active learning**, rater disagreement may be a significant information modality.

2.3 Related Problems in AI

We treat the subjectivity issue as being distinct from other cases that imply more than one label per sample, such as the presence of *hierarchical labels*, or applications like *acoustic event detection*, in which many events are present in one audio recording (possibly multiple times). The case of *multi-label classification* is related to a degree, if we assume that a data sample can have a soft label, i.e., a distribution over categorical labels. However, even in the soft label approach, we focus more on capturing the human subjectivity in labelling each sample.

3 Addressing Subjectivity

The issue we discuss here is **truth inference**, i.e., the extraction of a single hard label per sample, by treating the opinions of various raters as noisy observations of a true value.

3.1 Simple Fusion of Labels

Under the assumption that disagreeing, minority voices in annotation should be considered as mistakes, or random observation noise, we are naturally interested in getting a single, hard label per sample by performing a denoising fusion of the original labels, where each is given the same weight. For categorical labels, this amounts to majority voting, and for continuous ones to mean, or median averaging. A major motivation for adopting such an outlook, would be that the reduction to hard labels allows for the application of established AI literature on the application of interest.

This approach was used to leverage the utility of crowdsourcing, in order to cheaply annotate datasets, including many that became milestones in Deep Learning (DL) research, such as ImageNet [19], MS COCO [20], the MIT Places database [21], the SUN attribute dataset [9], and is still very popular [7,22].

3.2 Weighted Fusion of Labels

The output quality of annotators can vary widely due to differing levels of expertise [23] or personality characteristics, with some of them 'spamming' random answers [17]. As such, whereas the usage of crowdsourcing allows for fast and cheap solution to the annotation of large datasets, the observation noise that may be injected by the process introduces a requirement for increased quality control. The aim here is to estimate and assign trustworthiness values to each rater based on their performance with respect to the annotation task and/or other raters, such that their opinions are weighed differently.

The Evaluator Weighted Estimator (EWE) approach [24] is based on the calculation of rater-specific inter-reliability scores for weighing each rater's scores. EWE has successfully been used for improving the quality of affect recognition [25]. This approach is only meaningful if there is significant overlap among raters

with respect to samples, such that the aforementioned scores can be calculated with confidence. This is very often the case in data annotation in the laboratory setting, but not always; especially in the crowdsourcing case. In that case, the rater-specific scores cannot be based on inter-rater reliability, and instead can be calculated based on the performance consistency of each rater on the task, according to the Weighted Trustability Evaluator (WTE) [26] method.

The Case of Time-Continuous Labels. It is of great use to model affect in a continuous-valued, and continuous-time manner. When raters provide **sequences of continuous-valued emotion dimensions**, there is one more challenge to consider: **rater reaction lag**, which is often specific to the person. In these cases, the weighted fusion must be performed in a manner that accounts for lag-based discrepancies between raters as well.

In a study performed in [27], the authors proposed to weigh each rater based on an inter-rater correlation based score, as well as manually experimented with different lags per segment. In [28], a Canonical Correlation Analysis (CCA) approach is proposed in conjunction with time-warping on the latent space to accommodate lag discrepancies. Time warping with additional rank based annotations that reduce the subjectivity of continuous values was proposed in [29]. This issue has also been approached via Expectation-Maximisation (EM) [30], by assuming that the sequences provided by each rater are perturbed versions of a common ground truth. By assuming knowledge of which sequences were provided by which raters, reaction lag can be modelled specifically in a rater-specific manner [31]. Outside continuous affect recognition, a Bayesian dynamical model that models noisy crowdsourced time-series labels of financial data was proposed in [32].

4 Embracing Subjectivity

The main question now becomes whether it is preferable to model for a hypothetical "Average Jane", or whether we can approach the problem by explicitly taking into account the presence of very unique voices. This might mean that we assume that samples are inherently non-prototypical and ambiguous, or that we want to model individual raters to capture unique groups of thought, or even that rater disagreement is one more attribute of the data; to be learnt and predicted.

4.1 Incorporating Subjectivity in the Algorithm

As an intermediate step before fully embracing subjectivity as a source of opportunities instead of simply viewing it as a challenge to be solved before proceeding to the modelling stage, we now discuss the case according to which we indeed extract a single hard label per sample, as well as a measure of rater disagreement that is to be used as an additional input, target, or algorithm feature to improve learning.

Learning with priviledged information, or **master-class learning** [33] is a machine learning paradigm according to which each sample has additional information that facilitates learning during the training phase, but is not required for making predictions during testing. We consider the rater disagreement to be such priviledged information. In the study performed in [34], the sample annotation agreement (prototypicality in context), was used to weigh positively the loss of the corresponding samples, and in fact, samples that exhibited high disagreement were discarded. Similarly, less emphasis is placed in samples with less annotation confidence in the Gaussian Process (GP) based method proposed in [12].

The above methods assume that low rater agreement implies that the sample is inherently useless, and will hinder the training process. However, whereas the high disagreement implies a very subjective sample, deep learning is known to be robust to massive random observation noise [35], and perhaps simply downweighing, or removing such samples naively is not the best approach. To this end, a more appropriate solution might be **to learn what high rater disagreement means in the context of a particular dataset.** A multi-task framework was adopted by [13,36], in which the first task is the prediction of the fused hard emotion label, and the second is the prediction of the inter-rater disagreement of a sample. By adopting such a framework, the predictive performance of the first task is improved. Further to the subject of deciding which samples are more informative in the context of a dataset and task, the authors of [14] utilised, among others, annotator disagreement as an input to a teacher model that makes decisions on which samples are more informative towards training a base predictor.

Assuming that subjectivity is inherent in the application of interest, we might be interested in **being able to predict the inter-rater disagreement of a test sample**. This has been shown as a possibility [37], and has been used in the context of active learning [38] for emotion recognition. Such an approach was adopted in [3] in the digital health domain, for the identification of samples that would most benefit from a medical second opinion.

4.2 Assuming Soft Labels

Instead of fusing the differing labels into a single, hard label, one can calculate the empirical distribution thereof, define a soft label distribution per sample, and train a model on that. The soft label encodes the ambiguousness in ground truth for each sample, and allows for the model to learn label correlations as well, something that has been cited to be a significant contribution to the good performance of model distillation techniques [39]. This is an approach that has been adopted extensively, as in the studies performed in [16,40–42], and as part of [43]. In certain cases, samples for which a hard label cannot be extracted with majority voting due to a lack of a consensus are discarded, however the authors of [16] utilise these ambiguous samples in a soft label framework and observe a competitive performance using only the ambiguous data, and a clear improvement if all data are used.

It is also worth exploring the usefulness of inter-label correlations. The presence of multiple labels per sample allows for richer information available in the estimation of their positions in a low-dimensional manifold [44–46].

4.3 Explicitly Modelling a Rater

When we know the correspondence between labels and raters, we are given the opportunity to model the evaluation process of each particular rater, by utilising an ensemble of models. One of the motivations behind this approach is that in the case where one of the raters is an expert and the others are novices or spammers, the voice of the latter will outweigh that of the former regardless of whether we are using hard labels or soft.

In the methods proposed in [8,11,43,47–49] the authors estimate the model parameters, and a measure of rater trustworthiness in a joint manner. In cases where the number of raters is prohibitively high, there have been attempts to model schools of thought [18]. Using a separate model for explicitly modelling a rater has also been utilised for machine translation [50], and emotion recognition [51]. In the latter study, the model has a common base but multiple heads, each aimed at modelling a hard label as output by a different rater, and then a fusion is applied such that a soft label is predicted. A similar approach was used more recently, in a study about machine vision on biomedical images [52], in which the authors additionally propose that learning a sample-specific weighted averaging of raters' inputs, motivated by the fact that certain raters may be better at annotating certain types of data.

5 Active Learning Under Subjectivity

In the presence of rater disagreement, one possible avenue to improve the quality of the data would be to apply *repeated labelling*, i.e., request additional raters to label the data points, with the purpose of reducing the impact of each individual voice. In [41] it was shown that the simple strategy of relabelling all samples resulted to much better data quality, leading to better test performance; however, it was shown that by focusing the relabelling process on a small set of samples was a much better choice in the interest of budget constraints. The above technique is important to consider in cases where acquiring entirely new examples is more expensive than simply acquiring additional labels for existing ones.

Another related concept is that of **self-healing** [41,53]; in [53], repeated labelling is used to request additional samples to improve the label signal of a selected subset of the already labelled data. Self-healing is the adaptation of the labels of the rest according to the newly updated predictions of the classifier.

Even in the case where we need to label new samples through active learning, the rater agreement is important information to have; in [38], the authors have trained rater uncertainty models, and utilised the output value on unlabelled samples as a proxy of *data informativeness* for selecting samples to be annotated. Alternatively, in [54], the authors utilised the predictive uncertainty output of

Support Vector Machines (SVMs) to select both new examples and the number of raters required to achieve a desired level of agreement.

Most importantly, exploiting subjectivity and rater disagreement is an opportunity for the improvement of active learning. By using a GP that models each rater explicitly in [55], and by jointly modelling each rater's trustworthiness [56] better representation of uncertainty is achieved for improved active learning.

In certain cases, during the data annotation process, an **unsure option** was also provided to the raters for assigning to the most ambiguous examples, something that lightens their workload, as well. In [57], an active learning framework can learn on samples with unsure labels, albeit it does not make predictions that a sample is unsure.

Explicitly utilising explicit rater models that also estimate rater trustworthiness has been used successfully to improve active learning from crowds in [58].

In order to model complex, high dimensionality data, such as text, audio, images, and graphs, deep learning has been very successful; Bayesian deep learning [59] has allowed for the principled estimation of model parameters by defining a weight prior, as well as incorporating knowledge from the evidence in these domains. In [60], the authors use Bayesian deep active learning in the context of multiple annotators on an Amazon Alexa dateset.

6 Learning with Subjectivity as the Norm?

Despite the widespread proof of more sophisticated methods being better than unweighted label fusion, majority voting in order to extract a hard label is still being widely applied: for example, in the subjective application of discriminating between hate speech and simply offensive language on online social media [7] and even more recent work on the subject [61]. The output of the crowdsourcing workers in the context of [7] indicate that very often there is a confusion in the human perception between hate-speech and offensive language, indicative of an ambiguous task, or at the very least, multiple ambiguous samples.

We believe that there is great value in adopting subjectivity-aware AI methods as the norm, and discuss several possible frontiers, as well as related fields with which we believe a collusion would be profitable.

6.1 Subjectivity-Aware AI Pipeline

In order to fully embrace subjectivity in the AI pipeline, there is great value in adapting each stage such that it can accommodate it. Given our knowledge that rater performance can be variable [17] and task- and data- specific [62], expecting them to provide hard labels for samples they are unsure about, inevitably leads to lower data quality. The *unsure option* [57,63] for raters has been shown to be one possible addition to the annotation process that may provide the downstream stages with valuable **ambiguity ground-truth**.

Furthermore, by considering financial budgeting behind crowdsourcing, it might be possible to use a mixture of experts and novices. One more way to generate ground-truth ambiguity information is by using improved **interactive learning techniques** that allow for experts to evaluate a limited amount novice labels [64–66].

Recently, computational ways of modelling a rater's attention during the annotation process, for counteracting the **rater drift** problem have been applied to model annotation quality [62].

6.2 Uncertainty-Aware Deep Learning

There has been recent interest in quantifying predictive uncertainty using deep learning [59,67], not necessarily in the multiple annotator case. The authors of [59] propose a method that decomposes uncertainty of a test sample into two different factors, i.e., epistemic uncertainty that is due to lack of observed data at that area in data space, and aleatory uncertainty, that is representative of observation noise in labelling. They show how explicitly modelling for such uncertainty factors improves learning in both classification and regression computer vision tasks such as segmentation and depth prediction, and discuss the **explanatory capacities** of such an approach. The importance of, and an approach for deep uncertainty decomposition with explainability tangents have also been discussed in the digital health domain [68].

We believe that such methods can naturally be applied in the multiple annotator setting, and the decomposition of uncertainty can provide valuable insight towards understanding the degree to which a sample is mislabelled by certain raters, or whether it is inherently ambiguous, something that should have profound impact in the repeated labelling, and active learning from crowds domains. In fact, the authors of an earlier study in repeated labelling have made initial explorations towards using different definitions of uncertainty [41]. Keeping more recent developments in mind [59,68], an interesting question is: what is the relation between annotation subjectivity uncertainty and predictive aleatory uncertainty?

6.3 The Information Value of Data

It is important to keep in mind the assumptions made by adopting any of the aforementioned approaches. In certain approaches that use rater disagreement as priviledged information (see Subsect. 4.1), high disagreement is treated as an indicator for low sample quality, motivating the discarding of such data.

The relation between uncertainty and data informativeness is a decision that should be made in a **dataset- and task-dependent** manner [14,69,70], given that in certain cases, training is focused on hard samples in order to improve training [71], in others easy samples are utilised in the beginning of curriculum learning [72], and finally, in other cases, the middle-way is adopted [38].

The **quantification of the information value of data** is very impactful towards active learning [70], should be performed with labelling subjectivity in

mind [14], and should be performed in a dataset-specific mannner. For example, a framework for achieving such a joint quantification of sample value during active learning in an online manner through reinforcement learning has been proposed in [69].

6.4 Fairness in AI

It is important to develop AI frameworks that do not reinforce or reflect biases present in data. By modelling individual raters, or schools of thought (see Subsect. 4.3), greater capacity for capturing certain dimensions of bias is provided. We believe that a more thorough exploration of bias-aware methods [73] on subjective tasks is an avenue that will be explored to a great degree in the future.

7 Conclusions

Even though subjectivity is a well-known quality in certain applications and data and many approaches have been developed to address it, we feel that a paradigm shift towards treating it like an opportunity for improved modelling should be undertaken. We have summarised various groups of work that accommodate for the presence of multiple raters, based on their underlying philosophies, and have built upon them to incite discussion towards possible future opportunities.

References

1. Schuller, B.W.: Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. Commun. ACM **61**(5), 90–99 (2018)
2. Esteva, A., et al.: A guide to deep learning in healthcare. Nat. Med. **25**(1), 24–29 (2019)
3. Raghu, M., et al.: Direct uncertainty prediction for medical second opinions. In: Proceedings of the International Conference on Machine Learning, pp. 5281–5290 (2019)
4. Deutsch, F.M., LeBaron, D., Fryer, M.M.: What is in a smile? Psychol. Women Q. **11**(3), 341–352 (1987)
5. Fischer, A.H., Kret, M.E., Broekens, J.: Gender differences in emotion perception and self-reported emotional intelligence: a test of the emotion sensitivity hypothesis. PloS One **13**(1) (2018)
6. McCluskey, K.W., Albas, D.C.: Perception of the emotional content of speech by Canadian and Mexican children, adolescents, and adults. Int. J. Psychol. **16**(1–4), 119–132 (1981)
7. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the International AAAI Conference on Web and Social Media (2017)
8. Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., Krause, A.: Fake news detection in social networks via crowd signals. In: Companion Proceedings of the the Web Conference, pp. 517–524 (2018)
9. Patterson, G., Xu, C., Su, H., Hays, J.: The sun attribute database: beyond categories for deeper scene understanding. Int. J. Comput. Vis. **108**(1–2), 59–81 (2014)

10. Smyth, P., Fayyad, U.M., Burl, M.C., Perona, P., Baldi, P.: Inferring ground truth from subjective labelling of venus images. In: Proceedings of Advances in Neural Information Processing Systems, pp. 1085–1092 (1995)
11. Raykar, V.C., et al.: Learning from crowds. J. Mach. Learn. Res. **11**(Apr), 1297–1322 (2010)
12. Sharmanska, V., Hernández-Lobato, D., Miguel Hernandez-Lobato, J., Quadrianto, N.: Ambiguity helps: classification with disagreements in crowdsourced annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2194–2202 (2016)
13. Han, J., Zhang, Z., Schmitt, M., Pantic, M., Schuller, B.: From hard to soft: towards more human-like emotion recognition by modelling the perception uncertainty. In: Proceedings of the ACM International Conference on Multimedia, pp. 890–897. ACM (2017)
14. Rizos, G., Schuller, B.: Modelling sample informativeness for deep affective computing. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3482–3486. IEEE (2019)
15. Cowen, A.S., Keltner, D.: Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proc. Natl. Acad. Sci. **114**(38), E7900–E7909 (2017)
16. Ando, A., Kobashikawa, S., Kamiyama, H., Masumura, R., Ijima, Y., Aono, Y.: Soft-target training with ambiguous emotional utterances for DNN-based speech emotion classification. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 4964–4968. IEEE (2018)
17. Kazai, G., Kamps, J., Milic-Frayling, N.: Worker types and personality traits in crowdsourcing relevance labels. In: Proceedings of the ACM International Conference on Information and Knowledge Management, pp. 1941–1944 (2011)
18. Tian, Y., Zhu, J.: Learning from crowds in the presence of schools of thought. In: Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining, pp. 226–234 (2012)
19. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
20. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
21. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Proceedings of Advances in Neural Information Processing Systems, pp. 487–495 (2014)
22. Li, Y., Tao, J., Schuller, B., Shan, S., Jiang, D., Jia, J.: MEC 2016: the multimodal emotion recognition challenge of CCPR 2016. In: Tan, T., Li, X., Chen, X., Zhou, J., Yang, J., Cheng, H. (eds.) CCPR 2016. CCIS, vol. 663, pp. 667–678. Springer, Singapore (2016). https://doi.org/10.1007/978-981-10-3005-5_55
23. Zhang, C., Chaudhuri, K.: Active learning from weak and strong labelers. In: Proceedings of Advances in Neural Information Processing Systems, pp. 703–711 (2015)
24. Grimm, M., Kroschel, K.: Evaluation of natural emotions using self assessment manikins. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 381–385. IEEE (2005)
25. Schuller, B., Hantke, S., Weninger, F., Han, W., Zhang, Z., Narayanan, S.: Automatic recognition of emotion evoked by general sound events. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 341–344. IEEE (2012)

26. Hantke, S., Marchi, E., Schuller, B.: Introducing the weighted trustability evaluator for crowdsourcing exemplified by speaker likability classification. In: Proceedings of the International Conference on Language Resources and Evaluation, pp. 2156–2161 (2016)
27. Nicolaou, M.A., Gunes, H., Pantic, M.: Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Trans. Affect. Comput. **2**(2), 92–105 (2011)
28. Nicolaou, M.A., Pavlovic, V., Pantic, M.: Dynamic probabilistic CCA for analysis of affective behavior and fusion of continuous annotations. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1299–1311 (2014)
29. Booth, B.M., Mundnich, K., Narayanan, S.S.: A novel method for human bias correction of continuous-time annotations. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3091–3095. IEEE (2018)
30. Gupta, R., Audhkhasi, K., Jacokes, Z., Rozga, A., Narayanan, S.S.: Modeling multiple time series annotations as noisy distortions of the ground truth: an expectation-maximization approach. IEEE Trans. Affect. Comput. **9**(1), 76–89 (2016)
31. Mariooryad, S., Busso, C.: Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. IEEE Trans. Affect. Comput. **6**(2), 97–108 (2014)
32. Bakhtiari, B., Yazdi, H.S.: Bayesian filter based on the wisdom of crowds. Neurocomputing **283**, 181–195 (2018)
33. Vapnik, V., Izmailov, R.: Learning using privileged information: similarity control and knowledge transfer. J. Mach. Learn. Res. **16**(2023–2049), 2 (2015)
34. Kim, Y., Provost, E.M.: Leveraging inter-rater agreement for audio-visual emotion recognition. In: Proceedings of the International Conference on Affective Computing and Intelligent Interaction, pp. 553–559. IEEE (2015)
35. Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., Belongie, S.: Learning from noisy large-scale datasets with minimal supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 839–847 (2017)
36. Eyben, F., Wöllmer, M., Schuller, B.: A multitask approach to continuous five-dimensional affect sensing in natural speech. ACM Trans. Interact. Intell. Syst. **2**(1), 1–29 (2012)
37. Steidl, S., Batliner, A., Schuller, B., Seppi, D.: The hinterland of emotions: facing the open-microphone challenge. In: Proceedings of the International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1–8. IEEE (2009)
38. Zhang, Z., Deng, J., Marchi, E., Schuller, B.: Active learning by label uncertainty for acoustic emotion recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association (2013)
39. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
40. Jin, R., Ghahramani, Z.: Learning with multiple labels. In: Proceedings of Advances in Neural Information Processing Systems, pp. 921–928 (2003)
41. Ipeirotis, P.G., Provost, F., Sheng, V.S., Wang, J.: Repeated labeling using multiple noisy labelers. Data Min. Knowl. Disc. **28**(2), 402–441 (2014)
42. Kim, Y., Kim, J.: Human-like emotion recognition: multi-label learning from noisy labeled audio-visual expressive speech. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5104–5108. IEEE (2018)

43. Chou, H.-C., Lee, C.-C.: Every rating matters: joint learning of subjective labels and individual annotators for speech emotion classification. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5886–5890. IEEE (2019)
44. Zhang, H., Jiang, L., Xu, W.: Multiple noisy label distribution propagation for crowdsourcing. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 1473–1479. AAAI Press (2019)
45. Zhang, J., Sheng, V.S., Wu, J.: Crowdsourced label aggregation using bilayer collaborative clustering. IEEE Trans. Neural Netw. Learn. Syst. **30**(10), 3172–3185 (2019)
46. Liu, Y., Zhang, W., Yu, Y., et al.: Truth inference with a deep clustering-based aggregation model. IEEE Access **8**, 16 662–16 675 (2020)
47. Yan, Y., et al.: Modeling annotator expertise: learning when everybody knows a bit of something. In: Proceedings of the International Conference on Artificial Intelligence and Statistics, pp. 932–939 (2010)
48. Rodrigues, F., Pereira, F.C.: Deep learning from crowds. In: Proceedings of the AAAI Conference on Artificial Intelligence (2018)
49. Morales-Álvarez, P., Ruiz, P., Santos-Rodríguez, R., Molina, R., Katsaggelos, A.K.: Scalable and efficient learning from crowds with gaussian processes. Inf. Fusion **52**, 110–127 (2019)
50. Cohn, T., Specia, L.: Modelling annotator bias with multi-task Gaussian processes: an application to machine translation quality estimation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 32–42 (2013)
51. Fayek, H.M., Lech, M., Cavedon, L.: Modeling subjectiveness in emotion recognition with deep neural networks: ensembles vs soft labels. In: Proceedings of the International Joint Conference on Neural Networks, pp. 566–570. IEEE (2016)
52. Guan, M.Y., Gulshan, V., Dai, A.M., Hinton, G.E.: Who said what: modeling individual labelers improves classification. In: Proceedings of the AAAI Conference on Artificial Intelligence (2018)
53. Shu, Z., Sheng, V.S., Li, J.: Learning from crowds with active learning and self-healing. Neural Comput. Appl. **30**(9), 2883–2894 (2018)
54. Zhang, Y., Coutinho, E., Zhang, Z., Quan, C., Schuller, B.: Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions. In: Proceedings of the ACM International Conference on Multimodal Interaction, pp. 275–278 (2015)
55. Rodrigues, F., Pereira, F., Ribeiro, B.: Gaussian process classification and active learning with multiple annotators. In: Proceedings of the International Conference on Machine Learning, pp. 433–441 (2014)
56. Long, C., Hua, G.: Multi-class multi-annotator active learning with robust Gaussian process for visual recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2839–2847 (2015)
57. Zhong, J., Tang, K., Zhou, Z.-H.: Active learning from crowds with unsure option. In: Proceedings of the International Joint Conference on Artificial Intelligence (2015)
58. Calma, A., Sick, B.: Simulation of annotators for active learning: uncertain oracles. In: Proceedings of the ECML PKDD Interactive Adaptive Learning Workshop, p. 49 (2017)
59. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: Proceedings of Advances in Neural Information Processing Systems, pp. 5574–5584 (2017)

60. Yang, J., Drake, T., Damianou, A., Maarek, Y.: Leveraging crowdsourcing data for deep active learning an application: learning intents in alexa. In: Proceedings of the World Wide Web Conference, pp. 23–32 (2018)
61. Rizos, G., Hemker, K., Schuller, B.: Augment to prevent: short-text data augmentation in deep learning for hate-speech classification. In: Proceedings of the ACM International Conference on Information and Knowledge Management, pp. 991–1000 (2019)
62. Tu, J., Yu, G., Wang, J., Domeniconi, C., Zhang, X.: Attention-aware answers of the crowd. In: Proceedings of the 2020 SIAM International Conference on Data Mining, pp. 451–459. SIAM (2020)
63. Takeoka, K., Dong, Y., Oyamada, M.: Learning with unsure responses. In: Proceedings of the AAAI Conference on Artificial Intelligence. AAAI (2020)
64. Hu, Q., He, Q., Huang, H., Chiew, K., Liu, Z.: Learning from crowds under experts' supervision. In: Tseng, V.S., Ho, T.B., Zhou, Z.-H., Chen, A.L.P., Kao, H.-Y. (eds.) PAKDD 2014. LNCS (LNAI), vol. 8443, pp. 200–211. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-06608-0_17
65. Liu, M., Jiang, L., Liu, J., Wang, X., Zhu, J., Liu, S.: Improving learning-from-crowds through expert validation. In: Proceedings of the International Joint Conferences on Artificial Intelligence, pp. 2329–2336 (2017)
66. Liu, S., Chen, C., Lu, Y., Ouyang, F., Wang, B.: An interactive method to improve crowdsourced annotations. IEEE Trans. Vis. Comput. Graph. **25**(1), 235–245 (2018)
67. Rodrigues, F., Pereira, F.C.: Beyond expectation: deep joint mean and quantile regression for spatiotemporal problems. IEEE Trans. Neural Netw. Learn. Syst. (2020)
68. Kwon, Y., Won, J.-H., Kim, B.J., Paik, M.C.: Uncertainty quantification using bayesian neural networks in classification: application to biomedical image segmentation. Comput. Stat. Data Anal. **142**, 106816 (2020)
69. Haußmann, M., Hamprecht, F., Kandemir, M.: Deep active learning with adaptive acquisition. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 2470–2476. AAAI Press (2019)
70. Ghorbani, A., Zou, J.: Data shapley: equitable valuation of data for machine learning. In: Proceedings of the International Conference on Machine Learning, pp. 2242–2251 (2019)
71. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 761–769 (2016)
72. Pentina, A., Sharmanska, V., Lampert, C.H.: Curriculum learning of multiple tasks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5492–5500 (2015)
73. Kim, B., Kim, H., Kim, K., Kim, S., Kim, J.: Learning not to learn: training deep neural networks with biased data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9012–9020 (2019)