Modelling Variability in Human Annotator Simulation

Anonymous ACL submission

Abstract

Human annotator simulation (HAS) serves as a cost-effective substitute for human evaluation tasks such as data annotation and system assessment. It is important to incorporate the variability present in human evaluation into HAS, since it helps capture diverse subjective interpretations and mitigate potential biases and over-representation. This work introduces a novel framework for modelling variability in HAS. Conditional softmax flow (S-CNF) is proposed to model the distribution of subjective human annotations, which leverages diverse human annotations via meta-learning. This enables the efficient generation of annotations that exhibit human variability for unlabelled input. In addition, a wide range of evaluation metrics are adopted to assess the capability and efficiency of HAS systems in predicting the aggregated behaviours of human annotators, matching the distribution of human annotations, and simulating the inter-annotator disagreements. Results demonstrate that the proposed method achieves state-of-the-art performance on two real-world human evaluation tasks: emotion recognition and toxic speech detection[1].

1 Introduction

Human evaluation is fundamental to machine learning research. It guides processes such as data annotation and model assessment, including for instance perceptual quality evaluation (Ma et al., 2015; Talebi and Milanfar, 2018; Ramesh and Sanampudi, 2022), annotation generation for weak supervision (Ratner et al., 2016; Wu et al., 2022a), and model optimization based on human preference (Schatzmann et al., 2007; Gür et al., 2018; Ruiz et al., 2019; Shi et al., 2019; Lin et al., 2021). Collecting human annotations or evaluations often requires substantial resources and may expose human annotators to distressing and harmful content in sensitive tasks (*e.g.*, toxic speech detection, suicidal risk prediction, and depression detection). This inspires the exploration of human annotator simulation (HAS) as a scalable and cost-effective alternative, which facilitates large-scale dataset evaluation, benchmarking, and system comparisons.

Variability is a unique aspect of real-world human evaluation. Individual variations in cognitive biases, cultural backgrounds, and personal experiences (Hirschberg et al., 2003; Wiebe et al., 2004; Haselton et al., 2015) can lead to variability in human interpretation (Maniati et al., 2022). It has been argued that achieving a single deterministic "ground truth" in subjective tasks like human evaluation is not feasible, nor essential (Alm, 2011; Wu et al., 2022b). Therefore, HAS should incorporate the variability present in human evaluation rather than solely relying on majority opinions. This mitigates potential biases and over-representation in scenarios where dominant opinions could potentially overshadow minority viewpoints, thus promoting fairness and inclusivity.

This work investigates modelling subjective human annotation distributions and simulating human-like annotations. We propose a novel framework which formulates HAS as a zero-shot density estimation problem. A new model, conditional softmax flows (S-CNFs), is proposed which leverages diverse human annotations via meta-learning. This enables the efficient generation of annotations that exhibit human variability for unlabelled input. To the best of our knowledge, this is the first work that incorporates human variability into HAS without requiring human annotators to be dynamically involved in the process, while remaining scalable for large crowd-sourced datasets. Moreover, a range of evaluation metrics are adopted to assess the capability and efficiency of HAS systems regarding prediction of the majority opinion, estimation of the human annotation distribution, and simulation of the inter-annotator disagreements. The proposed

[1] Code will be available upon acceptance.

approach is evaluated on two real-world applications: emotion recognition and toxic speech detection. Results demonstrate that the proposed method achieves state-of-the-art performance on modelling variability in HAS.

2 Human Annotator Simulation (HAS)

2.1 The Variability in Human Evaluation is Valuable

Each individual's perception of the world is unique and influenced by their physical state and cognitive biases. This leads to diverse and subjective interpretations. Such subjectivity can be manifest in various tasks such as emotion recognition (Hirschberg et al., 2003; Mihalcea and Liu, 2006), perceptual quality assessment (Wiebe et al., 2004; Seshadrinathan et al., 2010), and user experience evaluation (Zen and Vanderdonckt, 2016). Rather than seeking to reduce the variability in annotations, it is important to account for annotators' subjective interpretations when designing a human annotator simulator. The importance of variability in HAS can be demonstrated by the following examples:

Revealing data ambiguity. Incorporating the variability in human perception empowers HAS to reveal potential ambiguity or complexity in data, providing valuable insights for further analysis.

Mitigating bias and over-representation. Incorporating the variability in human judgements prevents HAS from being biased towards a certain perspective and ignoring minority viewpoints, leading to a more inclusive representation of opinions where all viewpoints are given due consideration (Dixon et al., 2018; Hutchinson et al., 2020).

Improving model alignment. Optimization based on human feedback has led to superior performance on tasks such as text generation (Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023), which aligns the behaviour of language models with human preferences. HAS could be helpful in this task, as it is an efficient and cost-effective alternative to generating human feedback.

2.2 Problem Formulation

Denote an event as $\boldsymbol{d}_i$, which consists of a descriptor (*e.g.,* an utterance or text) $\boldsymbol{x}_i$ and a set of M_i human annotations $\mathcal{D}_i = \{\boldsymbol{\eta}_i^{(m)}\}_{m=1}^{M_i}$ for $\boldsymbol{x}_i$. Note that different events may be labelled by different sets of annotators. Given a dataset of training events $\mathcal{D} = \{(\boldsymbol{x}_i, \mathcal{D}_i)\}_{i=1}^{N}$, HAS aims to model the conditional annotation distribution $\mathrm{p}(\boldsymbol{\eta}_i|\boldsymbol{x}_i)$ given the observations $\mathcal{D}_i$ of $\boldsymbol{\eta}_i$ provided by different annotators. For a unseen test descriptor $\boldsymbol{x}_*$, HAS can then predict $\mathrm{p}(\boldsymbol{\eta}_*|\boldsymbol{x}_*)$ to simulate human-like annotations $\mathcal{D}_* = \{\boldsymbol{\eta}_*^{(m)}\}_{m=1}^{M_*}$ in a way that reflects how it would be labelled by human annotators.

2.3 Related Work

Prior work mainly investigated three approaches to simulating human annotations.

The first approach uses a single proxy variable $\boldsymbol{\eta}_i'$ (*e.g.*, majority vote) to summarize all annotations for each descriptor $\boldsymbol{x}_i$ (Kim et al., 2013; Djuric et al., 2015; Patton et al., 2016; Poria et al., 2017). This creates a proxy dataset $\mathcal{D}' = \{(\boldsymbol{x}_i, \boldsymbol{\eta}_i')\}_{i=1}^{N}$ and converts HAS into a supervised learning problem, which is usually solved by fitting a discriminative model to estimate the conditional distribution for the proxy variable. During testing, given an unseen descriptor $\boldsymbol{x}_*$, the model predicts the proxy variable $\boldsymbol{\eta}_*'$ for $\boldsymbol{x}_*$. Clearly, modelling a single proxy variable as in this approach fails to take into account the subjectivity and diversity in human behaviour and perception. Other work incorporated the variance of human annotations into the proxy variable (Deng et al., 2012; Prabhakaran et al., 2012; Plank et al., 2014; Dang et al., 2017; Han et al., 2017; Leng et al., 2021). However, all these approaches still focus on obtaining the "correct" label (*e.g.*, aiming for improved prediction accuracy) and minimizing the discrepancy among annotators (*e.g.*, reducing "noise" in annotations) rather than embracing inter-annotator disagreements.

The second approach explicitly models the behaviours of different annotators using different individual models in an ensemble or different heads in a single model (Fayek et al., 2016; Chou and Lee, 2019; Davani et al., 2022). This approach is computationally feasible only when the number of annotators is relatively small and when a sufficient quantity of annotation is available for each annotator, which is not applicable to large crowd-sourced datasets (Lotfian and Busso, 2019; Mathew et al., 2021) that are common in real-world applications.

The third approach approximates subjective probability distributions using Markov chain Monte Carlo with people (Sanborn and Griffiths, 2007; Harrison et al., 2020), which requires human annotators to be involved in the process in a dynamic setting. These methods present the descriptor $\boldsymbol{x}_*$ to human participants and asks them to provide a sequence of decisions $\mathcal{D}_*$ following

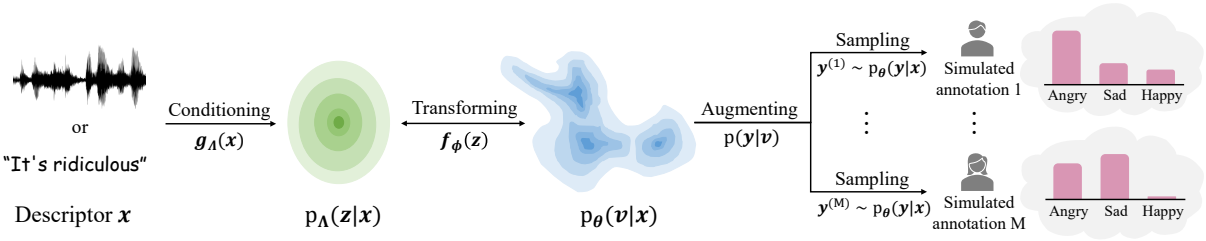

Figure 1: Diagram for the proposed zero-shot human annotator simulation framework.

the Metropolis-Hasting acceptance rule. The annotation distribution $\mathrm{p}(\boldsymbol{\eta}_*|\boldsymbol{x}_*)$ is then estimated based on $\mathcal{D}_*$. In other words, this requires access to human annotations $\mathcal{D}_*$ for estimating the annotation distribution for each $\boldsymbol{x}_*$, and there is no obvious way to transfer information between different events. Therefore, these methods cannot be applied to simulate annotation distributions for unlabelled test descriptors.

3 A Meta-learning Framework for HAS

This paper proposes a novel framework for HAS that meta-learns a conditional softmax flow (S-CNF) to estimate the human annotation distribution $\mathrm{p}(\boldsymbol{\eta}|\boldsymbol{x})$ across all training events $\mathcal{D}$. The proposed model learns to learn (*i.e.*, meta-learns) how to estimate the underlying distribution of human annotations $\mathcal{D}_i$ for any given descriptor $\boldsymbol{x}_i$ by leveraging the diverse human annotations, rather than designing a proxy variable to summarize $\mathcal{D}_i$ as in the first approach described in Sec. 2.3. Unlike the second approach in Sec. 2.3 which separately models each individual human annotator with a different model, our method is compatible with large crowd-sourced datasets since it amortizes across annotators with a single S-CNF model. Moreover, our model is a zero-shot human annotation simulator which can estimate the human annotation distribution $\mathrm{p}(\boldsymbol{\eta}_*|\boldsymbol{x}_*)$ for any unseen test descriptor $\boldsymbol{x}_*$ without access to any human annotations $\mathcal{D}_*$ for $\boldsymbol{x}_*$, in contrast to the third method in Sec. 2.3 which requires human annotators to be dynamically involved in the process of labelling $\boldsymbol{x}_*$.

3.1 A Latent Variable Model for HAS

The proposed framework for HAS is realized by a latent variable model[2]:

$$\mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) = \iint \mathrm{p}(\boldsymbol{y}|\boldsymbol{v})\mathrm{p}_{\boldsymbol{\phi}}(\boldsymbol{v}|\boldsymbol{z})\mathrm{p}_{\boldsymbol{\Lambda}}(\boldsymbol{z}|\boldsymbol{x})\mathrm{d}\boldsymbol{v}\mathrm{d}\boldsymbol{z}, \tag{1}$$

[2]For clarity, we use different notations for human annotations $\boldsymbol{\eta}$ and model outputs $\boldsymbol{y}$.

where the conditional prior $\mathrm{p}_{\boldsymbol{\Lambda}}(\boldsymbol{z}|\boldsymbol{x})$ learns to summarize useful information about the input descriptor $\boldsymbol{x}$ and encode the possible disagreements over $\boldsymbol{x}$ among different human annotators, which is helpful for the likelihood $\mathrm{p}_{\boldsymbol{\phi}}(\boldsymbol{y}|\boldsymbol{z}) = \int \mathrm{p}(\boldsymbol{y}|\boldsymbol{v})\mathrm{p}_{\boldsymbol{\phi}}(\boldsymbol{v}|\boldsymbol{z})\mathrm{d}\boldsymbol{v}$ to simulate human-like annotations for $\boldsymbol{x}$.

The proposed zero-shot human annotator simulator is illustrated in Figure 1. Specifically, the conditional prior is modelled by a conditional factorized Gaussian distribution $\mathrm{p}_{\boldsymbol{\Lambda}}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_{\boldsymbol{\Lambda}}(\boldsymbol{x}), \mathrm{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\Lambda}}(\boldsymbol{x})))$ whose mean $\boldsymbol{\mu}_{\boldsymbol{\Lambda}}(\boldsymbol{x})$ and variance $\boldsymbol{\sigma}^2_{\boldsymbol{\Lambda}}(\boldsymbol{x})$ are parameterized by a neural network $\boldsymbol{g}_{\boldsymbol{\Lambda}}$ with parameters $\boldsymbol{\Lambda}$.

The intermediate variable $\boldsymbol{v}$ is obtained by a deterministic invertible transformation $\mathrm{p}_{\boldsymbol{\phi}}(\boldsymbol{v}|\boldsymbol{z}) = \delta(\boldsymbol{v} - \boldsymbol{f}_{\boldsymbol{\phi}}(\boldsymbol{z}))$, where $\boldsymbol{f}_{\boldsymbol{\phi}}(\boldsymbol{z})$ is parameterized by an invertible neural network with parameters $\boldsymbol{\phi}$, and $\delta(\cdot)$ is the multivariate Dirac delta function. This results in a conditional normalizing flow (CNF):

$$\begin{aligned}\mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{x}) &= \int \delta(\boldsymbol{v} - \boldsymbol{f}_{\boldsymbol{\phi}}(\boldsymbol{z}))\mathrm{p}_{\boldsymbol{\Lambda}}(\boldsymbol{z}|\boldsymbol{x})\mathrm{d}\boldsymbol{z} \\ &= \mathrm{p}_{\boldsymbol{\Lambda}}\left(\boldsymbol{f}_{\boldsymbol{\phi}}^{-1}(\boldsymbol{v})\middle|\,\boldsymbol{x}\right)\left|\det\left(\frac{\partial \boldsymbol{f}_{\boldsymbol{\phi}}^{-1}(\boldsymbol{v})}{\partial \boldsymbol{v}}\right)\right|,\end{aligned} \tag{2}$$

where $\det(\cdot)$ denotes the determinant operator, $\partial \boldsymbol{f}_{\boldsymbol{\phi}}^{-1}(\boldsymbol{v})/\partial \boldsymbol{v}$ denotes the Jacobian matrix of $\boldsymbol{f}_{\boldsymbol{\phi}}^{-1}(\boldsymbol{v})$, and $\boldsymbol{\theta} := \{\boldsymbol{\phi}, \boldsymbol{\Lambda}\}$ denotes all parameters in this base CNF. This modelling choice has the advantage of having tractable marginal likelihood as in Eqn. (2) while not restricting the intermediate variable $\boldsymbol{v}$ to a specific type of distribution as in previous methods (*e.g.*, Gaussian (Han et al., 2017) and Student-t (Wu et al., 2023)), thus offering enhanced tractability, flexibility and generality. In addition, samples can be efficiently drawn from this model by first drawing $\boldsymbol{z} \sim \mathrm{p}_{\boldsymbol{\Lambda}}(\boldsymbol{z}|\boldsymbol{x})$ from the conditional prior and then computing the deterministic flow transformation $\boldsymbol{v} = \boldsymbol{f}_{\boldsymbol{\phi}}(\boldsymbol{z})$.

Finally, the output variable $\boldsymbol{y}$ is obtained by augmenting the intermediate variable $\boldsymbol{v}$ with the transformation $\mathrm{p}(\boldsymbol{y}|\boldsymbol{v})$. For continuous annotations, the identity transformation $\mathrm{p}(\boldsymbol{y}|\boldsymbol{v}) = \delta(\boldsymbol{y} - \boldsymbol{v})$ is used.

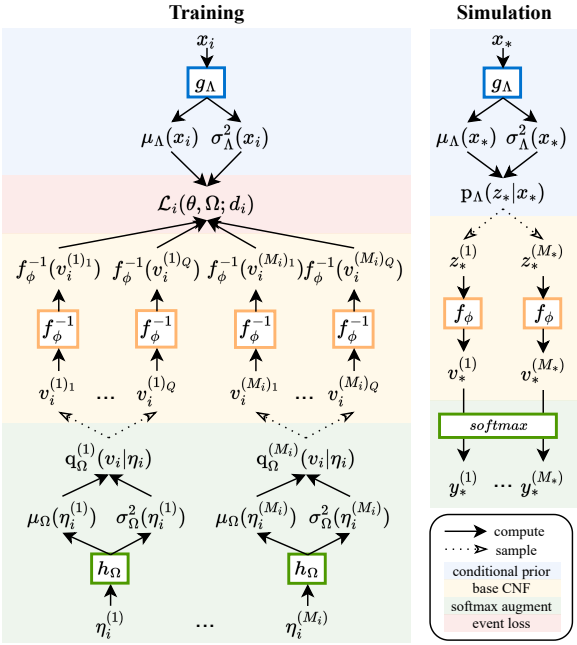

Figure 2: Illustration of the S-CNF workflow. $\boldsymbol{g_\Lambda}$ is the feature encoder, $\boldsymbol{f_\phi}$ is an invertible neural network, $\boldsymbol{h_\Omega}$ is the variational encoder.

However, real-world human evaluation tasks often involve discrete annotations (*e.g.,* human annotators are usually instructed to choose from a predefined set of options). In the next section, a new model class is introduced to accommodate discrete annotations.

3.2 Conditional Softmax Flows (S-CNFs)

We propose a new model class called conditional softmax flow (S-CNF) to accommodate discrete annotations. The workflow of S-CNF is illustrated in Figure 2. S-CNFs augment the base CNFs by applying the softmax function $\mathrm{p}(\boldsymbol{y}|\boldsymbol{v}) = \delta\left(\boldsymbol{y} - \mathrm{softmax}(\boldsymbol{v})\right)$ to transform the continuous intermediate variable $\boldsymbol{v}$ into categorical probabilities $\boldsymbol{y}$. Let c be a categorical variable with probability $\mathrm{P}(c = k|\boldsymbol{y}) = \boldsymbol{y}_k$ $(k = 1, \cdots, K)$, which represents the categorical annotation for an descriptor $\boldsymbol{x}$, with $\mathrm{P}(c = k|\boldsymbol{v}) = \int \boldsymbol{y}_k \delta\left(\boldsymbol{y} - \mathrm{softmax}(\boldsymbol{v})\right) \mathrm{d}\boldsymbol{y} = \mathrm{softmax}(\boldsymbol{v})_k$. The marginal likelihood of S-CNF is given by

$$\mathrm{P}_{\boldsymbol{\theta}}(c = k|\boldsymbol{x}) = \int \mathrm{P}(c = k|\boldsymbol{v})\mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{x})\mathrm{d}\boldsymbol{v} \qquad (3)$$

where $\mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{x})$ is the marginal likelihood of the base CNF defined in Eqn. (2). Since the marginal likelihood of the S-CNF given in Eqn. (3) is analytically intractable due to the softmax transformation, we propose to approximate it using variational inference (Wainwright et al., 2008) with a learnable mean-field Gaussian variational posterior $\mathrm{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{v}|\boldsymbol{\mu}_{\boldsymbol{\Omega}}(\boldsymbol{y}), \mathrm{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\Omega}}(\boldsymbol{y})))$, whose mean $\boldsymbol{\mu}_{\boldsymbol{\Omega}}(\boldsymbol{y})$ and variance $\boldsymbol{\sigma}^2_{\boldsymbol{\Omega}}(\boldsymbol{y})$ are parameterized by a neural network $\boldsymbol{h_\Omega}$ with parameters $\boldsymbol{\Omega}$. This can be seen as a probabilistic inverse of the softmax transformation $\mathrm{p}(\boldsymbol{y}|\boldsymbol{v})$. Applying Jensen's inequality to the log marginal likelihood of the S-CNF in Eqn. (3), we obtain a tractable evidence lower bound (ELBO):

$$\begin{aligned}\log \mathrm{P}_{\boldsymbol{\theta}}(c = k|\boldsymbol{x}) \geq \mathbb{E}_{\mathrm{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}|\boldsymbol{y})}[\log \mathrm{P}(c = k|\boldsymbol{v}) \\ + \log \mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{x}) - \log \mathrm{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}|\boldsymbol{y})].\end{aligned} \qquad (4)$$

It is worth noting that the softmax flow likelihood $\mathrm{P}(c = k|\boldsymbol{v}) = \mathrm{softmax}(\boldsymbol{v})_k$ places non-zero probability mass for every category $k = 1, \cdots, K$, which is different from argmax flow (Hoogeboom et al., 2021) whose likelihood only places probability mass for a single category. From a modelling perspective, softmax flow has a better capacity to represent the variability and uncertainty in human annotations. From an optimization perspective, the ELBO for softmax flow is always well-defined, whereas the ELBO for argmax flow is not defined when the model output does not match the human annotation in which case the log-likelihood would be $\log(0)$ and requires additional thresholding tricks to fix (Hoogeboom et al., 2021).

3.3 A Meta-learning Objective for S-CNFs

Using the variational approximation defined in Eqn. (4), the loss $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\Omega}; \boldsymbol{d}_i)$ for S-CNF on a single event $\boldsymbol{d}_i$ can be defined as the average negative ELBO evaluated on the set of the human annotations $\mathcal{D}_i = \{\boldsymbol{\eta}_i^{(m)}\}_{m=1}^{M_i}$ for the corresponding descriptor $\boldsymbol{x}_i$: $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\Omega}; \boldsymbol{d}_i) =$

$$\begin{aligned}-\frac{1}{M_i}\sum_{m=1}^{M_i} \mathbb{E}_{\mathrm{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}|\boldsymbol{\eta}_i^{(m)})}\left[\sum_{k=1}^{K} \boldsymbol{\eta}_{i,k}^{(m)} \log \mathrm{P}(c_i = k|\boldsymbol{v})\right. \\ \left. + \log \mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{x}_i) - \log \mathrm{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}|\boldsymbol{\eta}_i^{(m)})\right],\end{aligned} \qquad (5)$$

where the expectation over the variational posterior is approximated by Monte Carlo simulation with the reparameterization trick (Kingma and Welling, 2014). Following the episodic training scheme (Vinyals et al., 2016; Snell et al., 2017; Chen et al., 2023), we treat density estimation on each event as a learning problem and randomly sample a subset of such learning problems to train on at each step during meta-training. This results in a meta-learning objective across the training events in $\mathcal{D}$:

$$\mathcal{L}_{\mathrm{meta}}(\boldsymbol{\theta}, \boldsymbol{\Omega}; \mathcal{D}) = \mathbb{E}_{\boldsymbol{d}_i \sim \mathrm{p}(\mathcal{D})}[\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\Omega}; \boldsymbol{d}_i)], \qquad (6)$$

where $\mathrm{p}(\mathcal{D})$ denotes the uniform distribution over $\mathcal{D}$. Intuitively, this objective maps each human annotation to the latent space of the corresponding descriptor by the S-CNF during meta-training, which helps the model to build a diverse latent representation that captures the variability in human annotations across different descriptors.

At test time, the S-CNF can simulate human-like annotations for an unseen, unlabelled descriptor $\boldsymbol{x}_*$ by first drawing $\boldsymbol{v}_*^{(m)} \sim \mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{x}_*)$ from the base CNF then applying the softmax function $\boldsymbol{y}_*^{(m)} = \mathrm{softmax}(\boldsymbol{v}_*^{(m)})$ for $m = 1, \cdots, M_*$, where M_* denotes the number of annotations to be simulated. Note that each sample of S-CNF is a categorical distribution with probabilities $\boldsymbol{y}_*^{(m)}$. More details can be found in Appendix D.

4 Experimental Setup

The proposed framework for variability-aware HAS is evaluated on two real-world applications for speech and natural language processing: emotion class labelling and toxic speech detection. A wide range of evaluation metrics are adopted to assess the performance of HAS systems.

4.1 Evaluation Tasks and Datasets

Emotion class labelling. The highly subjective perception of emotion often results in disagreements among human annotators. Most emotion datasets employ multiple annotators to label each utterance while prior works typically use the majority vote as the ground-truth (Busso et al., 2008; Poria et al., 2019; Wu et al., 2021). A variability-aware HAS system better handles different opinions among human annotators and enhances the fairness of emotion class annotation. MSP-Podcast (Lotfian and Busso, 2019) is one of the largest publicly available datasets in speech emotion recognition, which contains natural English speech from podcast recordings annotated using crowd-sourcing. Each utterance has 6.7 annotations on average. Release 1.6 was used in our experiments, which contains 50k+ utterances from 1k+ speakers. The standard splits of training, validation and test were used. Emotion labels were grouped into five categories: angry, sad, happy, neutral, and other. 16.5% of the utterances do not have a majority agreed emotion class.

Toxic speech detection aims to filter out harmful language, which is crucial for respectful online environments and healthy communications among users. A variability-aware HAS system accounts for comprehensive understanding of hate speech, which is a good substitute for human annotators to reduce their exposure to distressing and harmful content. HateXplain (Mathew et al., 2021) was used in our experiments, which is a popular dataset for toxic speech detection, which contains 20k+ text posts from Twitter and Gab. These posts are labelled using crowd-sourcing with the commonly used three-category annotation: hate, offensive, normal. Each post is annotated by three annotators. Cases where all three annotators choose a different class (919 out of 20,148 posts) were originally excluded from the standard split of the dataset. We incorporate these cases into our training, validation, and test sets in an 8:1:1 ratio to better reflect the inter-annotator disagreements, resulting in 16,118 posts for training, 2,014 for validation, and 2,016 for testing.

4.2 Evaluation Metrics

A range of metrics are adopted to assess the empirical performance of HAS systems in terms of majority prediction, distribution matching, and human variability simulation.

Majority prediction. Classification accuracy (Acc) for the majority vote is evaluated for all test inputs with majority-agreed human annotations.

Distribution matching. Negative log likelihood (NLL) is used to evaluate how well the model estimates the human annotation distribution: $\mathrm{NLL}^{\mathrm{all}} = -\frac{1}{N}\sum_{i=1}^{N}(\frac{1}{M_i}\sum_{m=1}^{M_i}\log \mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i))$.

Inter-annotator disagreement simulation. Apart from evaluating the goodness of fit, additional metrics are adopted to explicitly measure how well the model simulates the variability and disagreements in human annotations: (i) the root mean squared error of the standard deviations of the annotations for all test inputs: $\mathrm{RMSE}^s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\sigma_i - s_i)^2}$, where $\sigma_i = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{M_i}\sum_{m=1}^{M_i}(\eta_{i,k}^{(m)} - \bar{\eta}_{i,k})^2}$ and $s_i = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{M_*}\sum_{m=1}^{M_*}(y_{i,k}^{(m)} - \bar{y}_{i,k})^2}$; (ii) the absolute error of the average standard deviations of the annotations for all test inputs: $\mathcal{E}(\bar{s}) = |\bar{\sigma} - \bar{s}|$, where $\bar{\sigma} = \sum_{i=1}^{N}\sigma_i$ and $\bar{s} = \sum_{i=1}^{N} s_i$; (iii) Fleiss's kappa ($\kappa$) (Fleiss, 1971), a real number between -1 and $+1$, with -1 indicating no observed agreement and $+1$ indicating perfect agreement. The absolute error between the kappas of human annotations (κ) and simulated annotations ($\hat{\kappa}$) for all test inputs is reported: $\mathcal{E}(\hat{\kappa}) = |\hat{\kappa} - \kappa|$.

4.3 Baselines

The proposed S-CNF method is compared to baselines of various types such as ensemble methods, Bayesian methods, and conditional generative models. This includes deep ensemble (**Ensemble**) (Lakshminarayanan et al., 2017), Monte-Carlo dropout (**MCDP**) (Gal and Ghahramani, 2016), Bayes-by-backprop (**BBB**) (Blundell et al., 2015), conditional variational autoencoder (**CVAE**) (Kingma and Welling, 2014), conditional argmax flow (**A-CNF**) (Hoogeboom et al., 2021), and Dirichlet prior network (**DPN**) (Malinin and Gales, 2018; Wu et al., 2022b). We fit each method to all available human annotations for all utterances in the training set, tune hyperparameters on the validation set, and report performance on the test set. $M_* = 100$ samples are used to compute evaluation metrics at test time. Ensemble only consists of 10 systems due to its expensive computational cost. Details about the configuration of the baseline models can be found in Appendix E.

4.4 Backbone Architecture

All compared methods use the same upstream-downstream feature encoder $\boldsymbol{g_\Lambda}$ to extract features from descriptors. The upstream model (Bommasani et al., 2021) is pre-trained on a large amount of unlabelled data to learn universal representations. WavLM (Chen et al., 2022) and RoBERTa (Liu et al., 2019) are used as the pre-trained upstream models for speech and text descriptors respectively. The downstream model consists of two Transformer encoder blocks followed by two fully connected (FC) layers, which are fine-tuned to target specific applications. The invertible flow model $\boldsymbol{f_\theta}$ uses three real NVP blocks (Dinh et al., 2017) and the variational encoder $\boldsymbol{h_\Omega}$ contains an FC layer. Details about the structure and implementation can be found in Appendix D and Appendix F.

5 Results and Analysis

The proposed method is evaluated according to the setup in Sec. 4. The evaluation results along with case study of representative examples demonstrates the superior capability of the proposed method in capturing the aggregated behaviours of human annotators, matching the distribution of human annotations, and simulating the variability of human interpretation. For each metric, mean value with standard error over three independent runs are reported for all methods.

Emotion	**Acc** (↑)	**NLL^{all}** (↓)
MCDP	0.582±0.003	1.423±0.012
Ensemble	**0.603±0.002**	1.458±0.004
BBB	0.565±0.010	1.459±0.011
DPN	0.581±0.006	1.459±0.011
CVAE	0.275±0.000	1.661±0.000
A-CNF	0.583±0.002	1.430±0.006
S-CNF	0.591±0.002	**1.403±0.011**
Toxic	**Acc** (↑)	**NLL^{all}** (↓)
MCDP	0.656±0.009	0.951±0.032
Ensemble	**0.682±0.002**	0.909±0.012
BBB	0.670±0.001	0.670±0.001
DPN	0.581±0.006	1.158±0.002
CVAE	0.406±0.000	1.150±0.000
A-CNF	0.628±0.003	0.892±0.011
S-CNF	0.673±0.002	**0.837±0.008**

Table 1: Comparison to the baselines in terms of majority prediction and distribution matching. CVAE collapses to one category for all inputs. The best value in each column is shown in bold and the second best is underlined.

5.1 Performance

Table 1 compares all methods in terms of majority prediction and distribution matching. The Ensemble achieves the best majority prediction accuracy (Acc) at the cost of training 10 independent systems. The proposed S-CNF achieves the second-best majority prediction accuracy with only a tenth of the computational cost of Ensemble during training and testing. Despite this, we stress that achieving the highest accuracy is not the goal of HAS since accuracy only measures the majority prediction performance and ignores variability in the annotations. More importantly, S-CNF is the best at matching the distributions of human annotations (in terms of NLL^{all}) among all compared methods. Table 2 reports the test results for all compared methods regarding inter-annotator disagreement annotator simulation. S-CNF again outperforms all compared methods in modelling the variability in human annotations, evident by the smallest RMSE^s, $\mathcal{E}(\bar{s})$ and $\mathcal{E}(\hat{\kappa})$.

5.2 Computational Time Cost

The computational time cost of all of compared methods for the two tasks studied in the paper are shown in Table 3. Denote M_* as the number of annotations to be simulated. The Ensemble model with M_* members involves training and testing M_* individual models, which costs $M_*\times$ more training

	Emotion class labelling			Toxic speech detection		
	RMSE^s (↓)	$\mathcal{E}(\bar{s})$ (↓)	$\mathcal{E}(\hat{\kappa})$ (↓)	RMSE^s (↓)	$\mathcal{E}(\bar{s})$ (↓)	$\mathcal{E}(\hat{\kappa})$ (↓)
MCDP	0.294±0.001	0.193±0.000	0.467±0.005	0.300±0.002	0.129±0.003	0.143±0.008
Ensemble	0.271±0.003	0.160±0.004	0.344±0.017	<u>0.289±0.001</u>	0.100±0.003	<u>0.064±0.006</u>
BBB	0.289±0.005	0.187±0.008	0.511±0.034	0.949±0.021	0.300±0.009	0.127±0.022
DPN	0.296±0.001	0.193±0.001	<u>0.104±0.016</u>	0.296±0.001	0.193±0.001	0.104±0.016
CVAE	0.333±0.000	0.244±0.000	—	0.345±0.000	0.208±0.000	—
A-CNF	<u>0.239±0.001</u>	<u>0.097±0.002</u>	0.382±0.015	0.297±0.001	<u>0.087±0.008</u>	0.198±0.027
S-CNF	**0.218±0.000**	**0.020±0.002**	**0.068±0.021**	**0.263±0.001**	**0.002±0.001**	**0.026±0.012**

Table 2: Comparison to the baselines in terms of inter-annotator agreement simulation. CVAE collapses to one category for all inputs. The best value in each column is shown in bold and the second best is underlined.

Emotion	Training (sec)	Testing (sec)
MCDP	7.20±0.10E+03	1.82±0.01E+04
Ensemble	1.46±0.00E+05	1.67±0.01E+03
BBB	7.55±0.01E+03	1.79±0.01E+04
DPN	6.80±0.01E+03	2.90±0.01E+02
A-CNF	7.04±0.02E+03	2.31±0.07E+02
S-CNF	6.99±0.00E+03	2.12±0.02E+02
Toxic	**Training (sec)**	**Testing (sec)**
MCDP	2.42±0.02E+02	5.99±0.02E+02
Ensemble	2.39±0.01E+03	4.00±0.04E+01
BBB	3.22±0.01E+02	5.79±0.01E+02
DPN	1.92±0.01E+02	2.67±0.02E+01
A-CNF	3.14±0.04E+02	1.40±0.11E+01
S-CNF	2.63±0.02E+02	1.37±0.09E+01

Table 3: Computational wall-clock time. The number M_* of annotations to simulate is set to 10 for ensemble and 100 for all other methods.

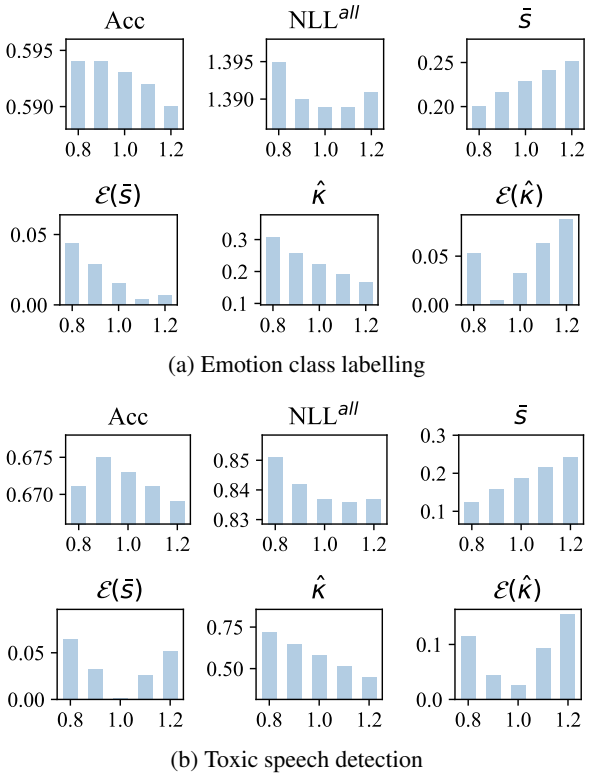

Figure 3: The effect of prior tempering on S-CNF. The x-axis corresponds to the prior temperature.

time and $M_*\times$ more testing time. MCDP and BBB require M_* forward passes during testing to generate M_* samples and therefore cost $M_*\times$ more testing time. All other methods require a single forward pass. S-CNF has slightly longer training time than DPN due to sampling procedure in ELBO while being the most efficient for testing.

5.3 Adjusting the Variability of S-CNFs by Prior Tempering

One advantage of S-CNF is that its sample variability can be easily controlled on demand without re-training by tempering the standard deviation of $\mathrm{p}_{\boldsymbol{\Lambda}}(\boldsymbol{z}|\boldsymbol{x})$ at test time. Figure 3 explores the effect of prior tempering on the performance. Overall, the trend is clear that the simulated annotations become more diverse as the temperature increases, shown by the increase in the average standard deviation ($\bar{s}$) and the decrease in Fleiss's kappa ($\hat{\kappa}$) of simulated annotations. As the temperature decreases, simulated annotations tends to concentrate more around the mean. The default temperature value 1 used during training (*i.e.*, no tempering) achieves the best trade-off among majority prediction accuracy (Acc), distribution matching (NLL^{all}), and inter-annotator disagreement simulation (in terms of $\mathcal{E}(\bar{s})$ and $\mathcal{E}(\hat{\kappa})$). In addition, prior tempering in S-CNF covers a wider range of dynamics than adjusting the dropout rate in MCDP. More details can be found in Appendix H.

5.4 Case Study

To better illustrate the properties of the annotations simulated by different methods, we visualize the

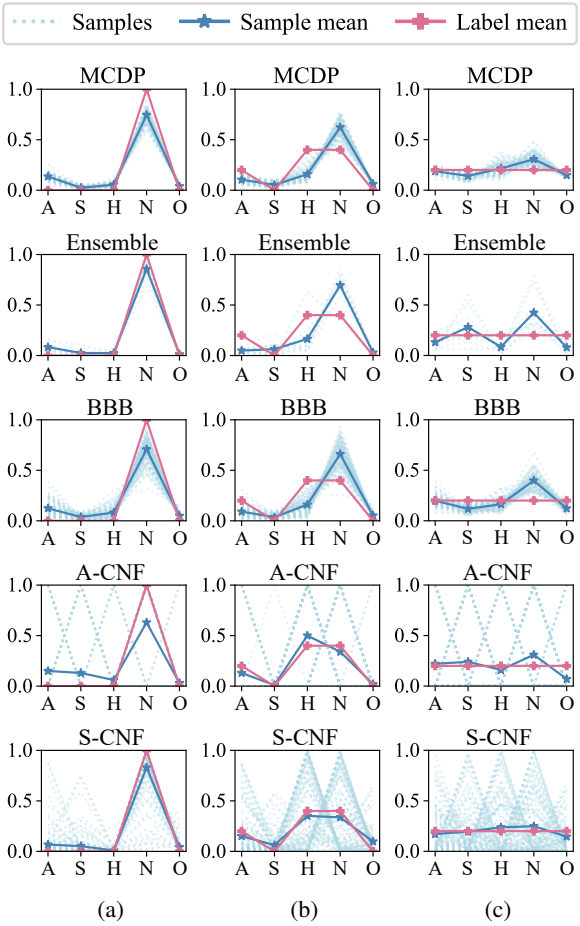

Figure 4: Visualization of simulated annotations on the emotion class labelling task for utterance (a) "0114_0263.wav", (b) '0167_0179_0001.wav", (c) "0574_0476.wav". The x-axis corresponds to emotion classes **A**ngry, **S**ad, **H**appy, **N**eutral, **O**thers. The y-axis corresponds to the probability mass. Each sample is a categorical distribution. The probability mass values of different categories in each categorical distribution are connected for the purpose of better visualization.

simulated distributions against the ground-truth distributions for three representative examples in Figure 4 (more case study examples can be found in Appendix I). Overall, the mean of the samples generated by S-CNF aligns the best with the average human label, indicating its superior performance in estimating the aggregated behaviours of human annotators. In addition, the samples generated by S-CNF are the most diverse among all compared methods, which manage to simulate the variability of the behaviours of different individual human annotators. In sharp contrast, the samples generated by all the other methods highly concentrate around their sample means. The visualized result for each example is analyzed below.

In case (a), human annotators reach a consensus. The majority of samples generated by S-CNF exhibit prominent peaks aligned with the ground-truth emotion class "neutral". In contrast, many samples generated by A-CNF peak at other emotion classes.

In case (b), human opinions diverge. The majority of samples generated by S-CNF are sharp categorical distributions peaking at one of the two majority emotion classes "happy" and "neutral". Additionally, a few samples generated by S-CNF peak at the emotion class "angry", which manages to simulate the minority viewpoint held by some annotators. Very few human annotators attribute this utterance to the emotion classes "sad" and "other", and S-CNF likewise produces scarce samples peaking at these classes.

In case (c), five human annotators give distinct emotion labels, resulting in a tie in the label means. The tie comes from annotators' diverse individual perceptions of the emotion rather than consensus on its ambiguity. S-CNF is the only model that can simulate both the diverse behaviours of different individual annotators and the aggregated behaviour of all annotators since the individual samples are sharp categorical distributions peaking at one of the five emotion classes and the mean of the samples aligns well with the label mean.

A case study of toxic speech detection exhibits similar trends and can be found in Appendix J.

6 Conclusions

This paper studied human annotator simulation (HAS), a cost-effective alternative to generating human-like annotations for automatic data labelling and model evaluation. A novel framework is proposed to incorporate the variability of human evaluations into HAS. It leverages diverse annotations to estimate the distribution of human annotations by meta-learning a conditional softmax flow (S-CNF) on large crowd-sourced datasets. This overcomes the drawbacks of prior work and enables efficient generation of annotations that exhibit human variability for unlabelled test inputs. The proposed method clearly and consistently outperformed a wide range of methods on emotion class labelling and toxic speech detection, achieving the best performance for human annotation distribution matching and inter-annotator disagreement simulation. It is hoped that the proposed method could help mitigate unfair biases and over-representation in HAS and reduce the exposure of human annotators to potentially harmful content, thus promoting ethical AI practices.

Limitations

This work focuses on categorical annotations, which is commonly used during human evaluation. Other types of annotations can be accommodated by designing suitable corresponding output transformations $\mathrm{p}(\boldsymbol{y}|\boldsymbol{v})$.

The proposed S-CNF is tested for two representative tasks: emotion recognition and toxic speech detection with speech and text as input respectively. We believe that the proposed method can also be general to other tasks, which is kept for future research directions.

Ethics Statement

In this work, all human annotations used for training were taken from existing publicly available corpora, and no new human annotations were collected.

It is hoped that this work could play a part in promoting ethical AI practice. Firstly, it has been shown that the proposed HAS system can capture the inherent variability in human judgements and help mitigate biases and the issue of over-representation, thus producing a more inclusive representation of human opinions. The proposed HAS system also has the potential to minimize human annotators' exposure to offensive and/or hateful content in some evaluation tasks such as HateXplain. However, as with most research in machine learning, new modelling techniques could be used by bad actors to cause harm more effectively, but we do not see how the proposed HAS system is more concerning than any other method in this regard.

References

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proc. ACL*, Portland, USA.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *Proc. ICML*, Lille, France.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E.M. Provost, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Wenlin Chen, Austin Tripp, and José Miguel Hernández-Lobato. 2023. Meta-learning adaptive deep kernel gaussian processes for molecular property prediction. In *Proc. ICLR*, Kigali, Rwanda.

Huang-Cheng Chou and Chi-Chun Lee. 2019. Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification. In *Proc. ICASSP*, Brighton, UK.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proc. NeurIPS*, Long Beach, USA.

Ting Dang, Vidhyasaharan Sethu, Julien Epps, and Eliathamby Ambikairajah. 2017. An investigation of emotion prediction uncertainty using gaussian mixture regression. In *Proc. Interspeech*, Stockholm, Sweden.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Jun Deng, Wenjing Han, and Björn Schuller. 2012. Confidence measures for speech emotion recognition: A start. In *Speech Communication; 10. ITG Symposium*, pages 1–4. VDE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL*, Minneapolis, USA.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2017. Density estimation using real NVP. In *Proc. ICLR*, Toulon, France.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proc. AAAI*, New Orleans, USA.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proc. WWW*, Florence, Italy.

H.M. Fayek, M. Lech, and L. Cavedon. 2016. Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels. In *Proc. IJCNN*, Vancouver, Canada.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proc. ICML*, New York, USA.

Izzeddin Gür, Dilek Hakkani-Tür, Gokhan Tür, and Pararth Shah. 2018. User modeling for task oriented dialogues. In *Proc. SLT*, Athens, Greece.

Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller. 2017. From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In *Proc. ACM MM*, Mountain View, USA.

Peter Harrison, Raja Marjieh, Federico Adolfi, Pol van Rijn, Manuel Anglada-Tort, Ofer Tchernichovski, Pauline Larrouy-Maestri, and Nori Jacoby. 2020. Gibbs sampling with people. In *Proc. NeurIPS*, Vancouver, Canada.

Martie G Haselton, Daniel Nettle, and Paul W Andrews. 2015. The evolution of cognitive bias. *The Handbook of Evolutionary Psychology*, pages 724–746.

Julia Hirschberg, Jackson Liscombe, and Jennifer Venditti. 2003. Experiments in emotional speech. In *Proc. SSPR*, Tokyo, Japan.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. In *Proc. NeurIPS*, Online.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proc. ACL*, Online.

Y. Kim, H. Lee, and E. M. Provost. 2013. Deep learning for robust feature generation in audiovisual emotion recognition. In *Proc. ICASSP*, Vancouver, Canada.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational Bayes. In *Proc. ICLR*, Banff, Canada.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proc. NeurIPS*, Long Beach, USA.

Yichong Leng, Xu Tan, Sheng Zhao, Frank Soong, Xiang-Yang Li, and Tao Qin. 2021. MBNet: MOS prediction for synthesized speech with mean-bias network. In *Proc. ICASSP*, Toronto, Canada.

Hsien-Chin Lin, Nurul Lubis, Songbo Hu, Carel van Niekerk, Christian Geishauser, Michael Heck, Shutong Feng, and Milica Gasic. 2021. Domain-independent user simulation with transformers for task-oriented dialogue systems. In *Proc. SIGDIAL*, Singapore City, Singapore.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

R. Lotfian and C. Busso. 2019. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

Kede Ma, Kai Zeng, and Zhou Wang. 2015. Perceptual quality assessment for multi-exposure image fusion. *IEEE Transactions on Image Processing*, 24(11):3345–3356.

Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. In *Proc. NeurIPS*, Montreal, Canada.

Georgia Maniati, Alexandra Vioni, Nikolaos Ellinas, Karolos Nikitaras, Konstantinos Klapsas, June Sig Sung, Gunu Jho, Aimilios Chalamandaris, and Pirros Tsiakoulis. 2022. SOMOS: The Samsung open MOS dataset for the evaluation of neural text-to-speech synthesis. In *Proc. Interspeech*, Incheon, Korea.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. In *Proc. AAAI*, Vancouver, Canada.

Rada Mihalcea and Hugo Liu. 2006. A corpus-based approach to finding happiness. In *Proc. AAAI Spring Symposium*, Stanford, USA.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Proc. NeurIPS*, New Orleans, USA.

Brian Patton, Yannis Agiomyrgiannakis, Michael Terry, Kevin Wilson, Rif A. Saurous, and D. Sculley. 2016. AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech. In *Proc, NeurIPS Workshop*, Barcelona, Spain.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proc. EACL*, Gothenburg, Sweden.

S. Poria, E. Cambria, R. Bajpai, and A. Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proc. ACL*, Florence, Italy.

Vinodkumar Prabhakaran, Michael Bloodgood, Mona Diab, Bonnie Dorr, Lori Levin, Christine D. Piatko, Owen Rambow, and Benjamin Van Durme. 2012. Statistical modality tagging from rule-based annotations and crowdsourcing. In *Proc. ExProM Workshop*, Jeju, Korea.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Alexander J. Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In *Proc. NeurIPS*, Barcelona, Spain.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Nataniel Ruiz, Samuel Schulter, and Manmohan Chandraker. 2019. Learning to simulate. In *Proc. ICLR*, New Orleans, USA.

Adam Sanborn and Thomas Griffiths. 2007. Markov chain Monte Carlo with people. In *Proc. NeurIPS*, volume Vancouver, Canada.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-based user simulation for bootstrapping a POMDP dialogue system. In *Proc. NAACL*, Vancouver, Canada.

Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K. Cormack. 2010. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, 19(6):1427–1441.

Weiyan Shi, Kun Qian, Xuewei Wang, and Zhou Yu. 2019. How to build user simulators to train RL-based dialog systems. In *Proc. EMNLP-IJCNLP*, Hong Kong, China.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Proc. NeurIPS*, Long Beach, USA.

Vincent Stimper, David Liu, Andrew Campbell, Vincent Berenz, Lukas Ryll, Bernhard Schölkopf, and José Miguel Hernández-Lobato. 2023. normflows: A PyTorch package for normalizing flows. *Journal of Open Source Software*, 8(86):5361.

Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *Proc. NeurIPS*, Barcelona, Spain.

Martin J Wainwright, Michael I Jordan, et al. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.

Renzhi Wu, Shen-En Chen, Jieyu Zhang, and Xu Chu. 2022a. Learning hyper label model for programmatic weak supervision. In *Proc. ICLR*, Online.

Wen Wu, Chao Zhang, and Philip Woodland. 2023. Estimating the uncertainty in emotion attributes using deep evidential regression. In *Proc. ACL*, Toronto, Canada.

Wen Wu, Chao Zhang, and Philip C. Woodland. 2021. Emotion recognition by fusing time synchronous and time asynchronous representations. In *Proc. ICASSP*, Toronto, Canada.

Wen Wu, Chao Zhang, Xixin Wu, and Philip Woodland. 2022b. Estimating the uncertainty in emotion class labels with utterance-specific Dirichlet priors. *IEEE Transactions on Affective Computing*.

Mathieu Zen and Jean Vanderdonckt. 2016. Assessing user interface aesthetics based on the intersubjectivity of judgment. In *Proc. BCS HCI*, Poole, UK.

A The Sources of Variability in Human Evaluation

Human perception refers to the process by which individuals interpret and make sense of the sensory information they receive from the environment. It involves the integration of sensory data, cognitive processes, emotions, and previous experiences. Subjective perception emphasizes that each individual's perception of the world is unique and influenced by their internal mental states, beliefs, attitudes, and past experiences. As a result, people can interpret and react to the same stimuli differently, leading to diverse and subjective perceptions.

Each person's sensory organs, such as eyes and ears, may have slight variations in sensitivity and acuity, leading to different perceptions of the same stimuli. Cognitive biases, the inherent mental shortcuts or tendencies that influence how humans perceive and process information, can lead to difference in judgment and decision-making. People's past experiences, cultural norms, and upbringing also shape their perceptions. Different cultural backgrounds can lead to distinct interpretations of the same event, leading to diverse reaction. The variability in humans can manifest in various tasks such as colour perception, emotion recognition, art appreciation, and feedback preferences.

Embracing and understanding the variability of human perception is vital for various research fields such as psychology, neuroscience, human-computer interaction, and so on, and has practical implications in designing human-centered systems and promoting empathy and diversity. It helps create products and interfaces that cater to diverse user needs and preferences in fields like human-computer interaction and user experience design. Being aware of the variability of perception is crucial in ethical decision-making. It help ensures that different perspectives and cultural sensitivities are considered, which helps identify and address potential biases that might disproportionately affect certain groups or lead to unfair outcomes.

B Derivations

Detailed derivations for the training objectives on a single dataset $\mathcal{D}_i = \{\boldsymbol{\eta}_i^{(1)}, \cdots, \boldsymbol{\eta}_i^{(M)}\}$ with $\boldsymbol{x}_i$ are presented in this section. For the simplicity of notations, the subscription i in our derivations will be omitted without ambiguity where possible. The meta-learning objectives presented in the paper are obtained by averaging such single-task objectives across tasks.

B.1 Objective Function for the Base CNF

Denote the empirical human annotation distribution as $\mathrm{p}_m(\boldsymbol{y}|\boldsymbol{x}) = \delta\left(\boldsymbol{y} - \boldsymbol{\eta}^{(m)}\right),\ m = 1, \cdots, M$ and model output distribution as $\mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$. The average KL divergence between them over all M human annotations for this input $\boldsymbol{x}$ is given by:

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}, \boldsymbol{x}) &= \frac{1}{M}\sum_{m=1}^{M} \mathcal{KL}\left(\mathrm{p}_m(\boldsymbol{y}|\boldsymbol{x}) \parallel \mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})\right) \\
&= \frac{1}{M}\sum_{m=1}^{M} \int \mathrm{p}_m(\boldsymbol{y}|\boldsymbol{x}) \log \frac{\mathrm{p}_m(\boldsymbol{y}|\boldsymbol{x})}{\mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})} \mathrm{d}\boldsymbol{y} \\
&= -\frac{1}{M}\sum_{m=1}^{M} \int \mathrm{p}_m(\boldsymbol{y}|\boldsymbol{x}) \log \mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) \mathrm{d}\boldsymbol{y} + \text{const} \\
&= -\frac{1}{M}\sum_{m=1}^{M} \log \mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{\eta}^{(m)}|\boldsymbol{x}) + \text{const}
\end{aligned}
$$

Minimizing this KL objective is equivalent to maximizing the average log likelihood $\log \mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{\eta}^{(m)}|\boldsymbol{x})$ over all human annotations as presented in the paper.

B.2 Objective Function for S-CNF

For categorical annotations, each label $\boldsymbol{\eta}^{(m)}$ represents the probabilities of all categories in the categorical human annotation distribution: $\boldsymbol{\eta}^{(m)} = [\boldsymbol{\eta}_1^{(m)}, \cdots, \boldsymbol{\eta}_K^{(m)}]$, where $\boldsymbol{\eta}_k^{(m)} = \mathrm{P}_m(c = k|\boldsymbol{x})$. Denote the model output distribution as $\mathrm{P}_{\boldsymbol{\theta}}(c|\boldsymbol{x})$. The average KL divergence between them over all M human annotations for this input $\boldsymbol{x}$ is given by:

$$
\begin{aligned}
\mathcal{L}^{\text{exact}}(\boldsymbol{\theta}; \mathcal{D}, \boldsymbol{x}) &= \frac{1}{M}\sum_{m=1}^{M} \mathcal{KL}\left(\mathrm{P}_m(c|\boldsymbol{x}) \parallel \mathrm{P}_{\boldsymbol{\theta}}(c|\boldsymbol{x})\right) \\
&= \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K} \mathrm{P}_m(c = k|\boldsymbol{x}) \log \frac{\mathrm{P}_m(c = k|\boldsymbol{x})}{\mathrm{P}_{\boldsymbol{\theta}}(c = k|\boldsymbol{x})} \\
&= -\frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K} \mathrm{P}_m(c = k|\boldsymbol{x}) \log \mathrm{P}_{\boldsymbol{\theta}}(c = k|\boldsymbol{x}) \\
&\quad + \text{const} \\
&= -\frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K} \boldsymbol{\eta}_k^{(m)} \log \mathrm{P}_{\boldsymbol{\theta}}(c = k|\boldsymbol{x}) + \text{const},
\end{aligned}
$$

where the marginal likelihood is lower bounded using variational inference:

$$\begin{aligned}
&\log \mathrm{P}_{\boldsymbol{\theta}}(c=k|\boldsymbol{x}) = \log \int \mathrm{P}(c=k|\boldsymbol{v})\mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{x})\mathrm{d}\boldsymbol{v} \\
&= \log \int \boldsymbol{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}|\boldsymbol{\eta}) \frac{\mathrm{P}(c=k|\boldsymbol{v})\mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{x})}{\boldsymbol{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}|\boldsymbol{\eta})}\mathrm{d}\boldsymbol{v} \\
&\geq \int \boldsymbol{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}|\boldsymbol{\eta}) \log \frac{\mathrm{P}(c=k|\boldsymbol{v})\mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{x})}{\boldsymbol{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}|\boldsymbol{\eta})}\mathrm{d}\boldsymbol{v} \\
&= \mathbb{E}_{\boldsymbol{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}|\boldsymbol{\eta})}\left[\log \mathrm{P}(c=k|\boldsymbol{v}) + \log \mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{x}) \right. \\
&\qquad \left. - \log \boldsymbol{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}|\boldsymbol{\eta})\right].
\end{aligned}$$

Therefore, the final negative ELBO objective is obtained by

$$\begin{aligned}
&\mathcal{L}^{\text{exact}} = -\frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K}\boldsymbol{\eta}_k^{(m)} \log \mathrm{P}_{\boldsymbol{\theta}}(c=k|\boldsymbol{x}) \\
&\leq -\frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K}\boldsymbol{\eta}_k^{(m)}\mathbb{E}_{\boldsymbol{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}|\boldsymbol{\eta}^{(m)})}\left[\log \mathrm{P}(c=k|\boldsymbol{v}) \right. \\
&\qquad \left. + \log \mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{x}) - \log \boldsymbol{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}|\boldsymbol{\eta}^{(m)})\right] \\
&= -\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\boldsymbol{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}|\boldsymbol{\eta}^{(m)})}\left[\sum_{k=1}^{K}\boldsymbol{\eta}_k^{(m)} \log \mathrm{P}(c=k|\boldsymbol{v})) \right. \\
&\qquad \left. + \log \mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{x}) - \log \boldsymbol{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}|\boldsymbol{\eta}^{(m)})\right] \\
&= \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \mathcal{D}, \boldsymbol{x}),
\end{aligned}$$

where

$$\begin{aligned}
&\log \mathrm{P}(c=k|\boldsymbol{v}) = \mathrm{logsoftmax}(\boldsymbol{v})_k, \\
&\log \mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{x}) = \mathrm{p}_{\boldsymbol{\Lambda}}\left(\boldsymbol{f}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{v})|\boldsymbol{x}\right)\left|\det\left(\frac{\partial \boldsymbol{f}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{v})}{\partial \boldsymbol{v}}\right)\right|, \\
&\log \boldsymbol{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}|\boldsymbol{\eta}^{(m)}) = \mathcal{N}(\boldsymbol{v}|\boldsymbol{\mu}_{\boldsymbol{\Omega}}(\boldsymbol{\eta}^{(m)}), \mathrm{diag}(\boldsymbol{\sigma}_{\boldsymbol{\Omega}}^2(\boldsymbol{\eta}^{(m)}))).
\end{aligned}$$

B.3 The Negative Log Likelihood ($\mathrm{NLL}_i^{\text{all}}$) for Categorical Annotations

The marginal likelihood of S-CNF is intractable, which can be approximated using Monte-Carlo simulation:

$$\begin{aligned}
&\mathrm{P}_{\boldsymbol{\theta}}(c=k|\boldsymbol{x}) = \int \mathrm{P}(c=k|\boldsymbol{v})\mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{x})\mathrm{d}\boldsymbol{v} \\
&= \mathbb{E}_{\mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{x})}\left[\mathrm{P}(c=k|\boldsymbol{v})\right] \\
&\approx \frac{1}{Q}\sum_{j=1}^{Q}\mathrm{P}(c=k|\boldsymbol{v}_j), \quad \{\boldsymbol{v}_j\}_{j=1}^{Q} \sim_{\text{iid}} \mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{x}) \\
&= \frac{1}{Q}\sum_{j=1}^{Q}\mathrm{softmax}(\boldsymbol{v}_j)_k, \quad \{\boldsymbol{v}_j\}_{j=1}^{Q} \sim_{\text{iid}} \mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{x}) \\
&= \bar{\boldsymbol{y}}_k,
\end{aligned}$$

where $\bar{\boldsymbol{y}} = \frac{1}{Q}\sum_{j=1}^{Q}\mathrm{softmax}(\boldsymbol{v}_j) = \frac{1}{Q}\sum_{j=1}^{Q}\boldsymbol{y}_j$ which is the average of the simulated categorical distributions. Let $\bar{\boldsymbol{\eta}} = \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\eta}^{(m)}$ be the average label.

Then, the $\mathrm{NLL}_i^{\text{all}}$ for a single input $\boldsymbol{x}_i$ is given by

$$\begin{aligned}
\mathrm{NLL}_i^{\text{all}} &= -\frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K}\boldsymbol{\eta}_{i,k}^{(m)} \log \mathrm{P}_{\boldsymbol{\theta}}(c=k|\boldsymbol{x}_i) \\
&\approx -\frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K}\boldsymbol{\eta}_{i,k}^{(m)} \log \bar{\boldsymbol{y}}_{i,k} \\
&= -\sum_{k=1}^{K}\bar{\boldsymbol{\eta}}_{i,k} \log \bar{\boldsymbol{y}}_{i,k},
\end{aligned}$$

which is the cross entropy between the averaged label and averaged sample.

C Emotion Label Processing for MSP-Podcast

In MSP-Podcast, each annotator can choose from ten emotion classes to label the primary emotion of an utterance: *Angry, Sad, Happy, Surprise, Fear, Disgust, Contempt, Neutral, Other*. Although only one option is allowed, they can say *other* and define their own emotion class which can be more than one. During label processing, the original *other* class is split into sub-classes depending on the manual defined label and merged with the predefined labels. The grouping details are shown as follows: (i) *Angry* includes *angry, disgust, contempt, annoyed*; (ii) *Sad* includes *sad, frustrated, disappointed, depressed, concerned*; (iii) *Happy* includes *happy, excited, amused*; (iv) *Neutral* includes *neutral*; (v) *Other* includes all other emotion subclasses not listed above.

D Model Structure Details

The procedure of sampling from and optimizing S-CNF are summarized in Algorithm 1 and 2.

A neural-network-based encoder $\boldsymbol{g}_{\boldsymbol{\Lambda}}$ is built to model $\boldsymbol{\mu}_{\boldsymbol{\Lambda}}(\boldsymbol{x}), \boldsymbol{\sigma}_{\boldsymbol{\Lambda}}^2(\boldsymbol{x})$ given input $\boldsymbol{x}$ where $\boldsymbol{\Lambda}$ is the model parameters. $\boldsymbol{g}_{\boldsymbol{\Lambda}}$ follows an upstream-downstream paradigm. The upstream model is pre-trained on large amount of unlabelled data to learn universal representations. The downstream model uses the learned representation from the upstream model for specific applications.

For tasks involving speech as input (*i.e.*, emotion class labelling), WavLM (Chen et al., 2022) is used

Algorithm 1 Sampling from S-CNF

Input: $\boldsymbol{x}$
Output: Categorical probability $\boldsymbol{y}$
Compute $\boldsymbol{\mu}_{\boldsymbol{\Lambda}}(\boldsymbol{x}), \boldsymbol{\sigma}^2_{\boldsymbol{\Lambda}}(\boldsymbol{x}) = \boldsymbol{g}_{\boldsymbol{\Lambda}}(\boldsymbol{x})$
Sample $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\Lambda}}(\boldsymbol{x}), \mathrm{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\Lambda}}(\boldsymbol{x}))$
Compute $\boldsymbol{v} = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{z})$
Compute $\boldsymbol{y} = \mathrm{softmax}(\boldsymbol{v})$

Algorithm 2 Optimizing S-CNF

Input: $\boldsymbol{x}, \mathcal{D} = \{\boldsymbol{\eta}^{(1)}, \cdots, \boldsymbol{\eta}^{(M)}\}$
Output: ELBO $\mathcal{L}^{\mathrm{ELBO}}$ on dataset $\mathcal{D}$
for $m = 1, \cdots, M$ **do**
  Compute $\boldsymbol{\mu}_{\boldsymbol{\Omega}}(\boldsymbol{\eta}^{(m)}), \boldsymbol{\sigma}^2_{\boldsymbol{\Omega}}(\boldsymbol{\eta}^{(m)}) = \boldsymbol{h}_{\boldsymbol{\Omega}}(\boldsymbol{\eta}^{(m)})$
  for $j = 1, \cdots, Q$ **do**
    Sample $\boldsymbol{v}_j \sim \boldsymbol{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}|\boldsymbol{\eta}^{(m)})$
    Compute $\mathcal{L}_j^{(m)} = -\sum_{k=1}^{K} \boldsymbol{\eta}_k^{(m)} \log \mathrm{P}(c = k|\boldsymbol{v}_j) + \log \mathrm{p}_{\boldsymbol{\theta}}(\boldsymbol{v}_j|\boldsymbol{x}) - \log \boldsymbol{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}_j|\boldsymbol{\eta}^{(m)})$
  end for
  Compute $\mathcal{L}_m^{\mathrm{ELBO}} = \frac{1}{Q}\sum_{j=1}^{Q} \mathcal{L}_j^{(m)}$
end for
Compute $\mathcal{L}^{\mathrm{ELBO}} = \frac{1}{M}\sum_{m=1}^{M} \mathcal{L}_m^{\mathrm{ELBO}}$

as the upstream model. WavLM is a speech foundation model pre-trained by self-supervised learning that takes raw waveform as input. The waveform is encoded by a CNN encoder followed by multiple Transformer encoders. The BASE+ version[3] of the model is used in this work which has 12 Transformer encoder blocks with 768-dimensional hidden states and 8 attention heads. The parameters of the pretrained WavLM are frozen and the weighted sum of the outputs of the 12 Transformer encoder blocks is used as the speech embeddings feeding into the downstream model.

RoBERTa (Liu et al., 2019) is used as upstream model to encode text input for toxic speech detection, which is a robustly optimized model of BERT (Devlin et al., 2019). RoBERTa is a Transformer-based language model pretrained on a large corpus of English data with the masked language modelling objective. The BASE version[4] was used in the work which has 12 Transformer layers, 768 hidden units, 12 attention heads, and 125 million parameters.

The downstream model consists of two Transformer encoder blocks followed by two FC layers. The Transformer encoder layers has a hidden dimension of 128 and four attention head. The output layer contains two heads to predict the mean and standard deviation of the latent distribution $\mathrm{p}_{\boldsymbol{\Lambda}}(\boldsymbol{z}|\boldsymbol{x})$. The invertible flow model $\boldsymbol{f}_{\boldsymbol{\theta}}$ uses three real NVP block (Dinh et al., 2017) of dimension 64. The variational encoder $\boldsymbol{h}_{\boldsymbol{\Omega}}$ for S-CNF contains a FC layer of dimension 64 and two output heads for the mean and standard deviation of the variational distribution $\boldsymbol{q}_{\boldsymbol{\Omega}}(\boldsymbol{v}|\boldsymbol{y})$.

[3]https://huggingface.co/microsoft/wavlm-base-plus
[4]https://huggingface.co/roberta-base

E Detailed Configuration of All Compared Methods

Ensemble consists of 10 systems initialized and trained using different random seeds. MCDP uses dropout rate of 0.4. A standard Gaussian prior is used for BBB. A modified version of EDL is used (Wu et al., 2023) which is trained by maximizing the per-observation-based marginal likelihood with a modified regularization term. Ensemble, MCDP, BBB, EDL use the same model structure as $\boldsymbol{g}_{\boldsymbol{\Lambda}}$ apart from removing the output head for predicting variance of latent distribution. A modified version of DPN (Wu et al., 2022b) is used which is trained by interpolating per-observation-based marginal likelihood with KL divergence. The coefficient of KL term is set to 5.0 for emotion class labelling and 2.0 for toxic speech detection. CVAE has the same $\boldsymbol{g}_{\boldsymbol{\Lambda}}$ structure as S-CNF for modelling $\mathrm{p}(\boldsymbol{z}|\boldsymbol{x})$, and two 64-d FC layers are used for encoder and decoder. A-CNF has identical model structure as S-CNF.

F Implementation Details

The system was implemented using PyTorch with the SpeechBrain (Ravanelli et al., 2021) and normflows (Stimper et al., 2023) toolkit. The Adadelta optimizer was used with an initial learning rate of 1.2 for emotion class labelling and 0.05 for speech quality assessment. The NewBob learning rate scheduler was used with annealing factor 0.8 and patience 1. The system was trained for 30 epochs and the model with the best validation performance was used for testing. The number of ELBO samples was set to 20. All experiments were run with three different seeds and the average and standard error are reported.

Emotion class labelling			
	RMSE^s	**MAE^s**	$\mathcal{E}(\bar{s})$
MCDP	0.305	0.233	0.206
Ensemble	0.277	0.222	0.166
BBB	0.284	0.226	0.178
CVAE	0.333	0.244	0.244
DPN	0.297	0.236	0.191
A-CNF	0.223	0.209	0.046
S-CNF	0.218	0.198	0.015
Toxic speech detection			
	RMSE^s	**MAE^s**	$\mathcal{E}(\bar{s})$
MCDP	0.297	0.242	0.122
Ensemble	0.290	0.220	0.105
BBB	0.279	0.229	0.115
CVAE	0.345	0.208	0.208
DPN	0.299	0.220	0.178
A-CNF	0.274	0.232	0.062
S-CNF	0.263	0.206	0.002

Table 4: Analysis of standard deviation of simulated samples.

G Analysis of Standard Deviation of Simulated Samples

It has been observed in Sec. 5.1 that flow models tend to have a larger difference between RMSE^s and $\mathcal{E}(\bar{s})$. This section provides detailed analysis to this observation. Let N be the number of test utterances. Three std-related metrics are computed: (i) RMSE between std of predictions and human labels: RMSE^s $= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_i - \sigma_i)^2}$; (ii) Mean absolute error between std of predictions and std of human labels: MAE^s $= \frac{1}{N}\sum_{i=1}^{N}|s_i - \sigma_i|$; (iii) Absolute error between average std of predictions and average std of human labels $\mathcal{E}(\bar{s}) = |\bar{s}_i - \bar{\sigma}_i|$. Results are shown in Table 4. The flow model tends to have larger discrepancy between MAE^s and $\mathcal{E}(\bar{s})$. According to the triangular inequality:

$$\mathcal{E}(\bar{s}) = \left|\frac{1}{N}\sum_{i=1}^{N} s_i - \frac{1}{N}\sum_{i=1}^{N}\sigma_i\right|$$
$$= \left|\frac{1}{N}\sum_{i=1}^{N}(s_i - \sigma_i)\right| \leq \frac{1}{N}\sum_{i=1}^{N}|s_i - \sigma_i| = \text{MAE}^s$$

which show that $\mathcal{E}(\bar{s})$ is a lower bound of MAE^s. The equality condition is satisfied when all samples are uniformly either greater than or less than the compared value. Therefore, a larger discrepancy between these two values indicates that the standard deviation of some samples exceeds that of the labels, while for others, it is lower. A smaller discrepancy indicates that the standard deviation of samples tend to be consistently larger of smaller than that of the labels. In Figure 5, 100 test utterances are randomly selected and the std of samples generated by different models are plotted, which supports the above conclusion. The proposed S-CNF has the best performance for matching the diversity of human annotations.

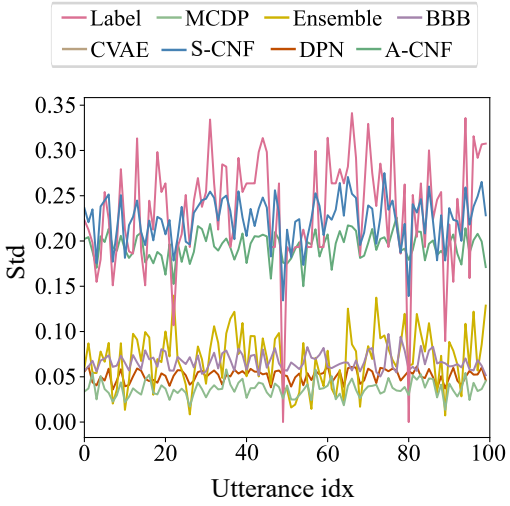

Figure 5: Standard deviation of simulated samples for emotion class labelling.

H Adjusting the Variability of CNFs by Prior Tempering

Sec. 5.3 has explored the effect of prior tempering on the performance. More details are provided in this section. Table 5 shows the effect of prior tempering on the performance of S-CNF. Table 6 shows the effect of dropout rate on the performance of MCDP. When the temperature increases or dropout rate increases, the simulated annotations become more diverse, shown by the increase in the average standard deviation ($\bar{s}$) and the decrease in Fleiss's kappa ($\hat{\kappa}$) of simulated annotations. Comparing two tables, it can be seen that prior tempering in CNF is more efficient and covers a wider range of dynamics than adjusting the dropout rate in MCDP.

Emotion class labelling							
T	**Acc**	**NLL^{all}**	**RMSE^{s}**	$\bar{s}$	$\mathcal{E}(\bar{s})$	$\hat{\kappa}$	$\mathcal{E}(\hat{\kappa})$
0.8	0.594	1.395	0.221	0.200	0.044	0.307	0.053
0.9	0.594	1.390	0.219	0.216	0.029	0.259	0.005
1.0	0.593	1.389	0.218	0.229	0.015	0.222	0.032
1.1	0.592	1.389	0.218	0.241	0.004	0.191	0.063
1.2	0.590	1.391	0.219	0.251	0.007	0.166	0.088
Toxic speech detection							
T	**Acc**	**NLL^{all}**	**RMSE^{s}**	$\bar{s}$	$\mathcal{E}(\bar{s})$	$\hat{\kappa}$	$\mathcal{E}(\hat{\kappa})$
0.8	0.671	0.851	0.272	0.125	0.065	0.721	0.115
0.9	0.675	0.842	0.265	0.157	0.033	0.650	0.044
1.0	0.673	0.837	0.263	0.188	0.002	0.580	0.026
1.1	0.671	0.836	0.264	0.216	0.026	0.512	0.094
1.2	0.669	0.837	0.267	0.242	0.052	0.450	0.156

Table 5: Adjusting the variability of CNFs by prior tempering.

Emotion class labelling							
dp	**Acc**	**NLL^{all}**	**RMSE^{s}**	$\bar{s}$	$\mathcal{E}(\bar{s})$	$\hat{\kappa}$	$\mathcal{E}(\hat{\kappa})$
0.1	0.583	1.463	0.303	0.040	0.205	0.791	0.537
0.2	0.589	1.426	0.303	0.040	0.204	0.773	0.519
0.3	0.590	1.415	0.300	0.045	0.199	0.761	0.507
0.4	0.585	1.405	0.296	0.051	0.194	0.723	0.469
0.5	0.589	1.409	0.294	0.053	0.191	0.715	0.461
Toxic speech detection							
dp	**Acc**	**NLL^{all}**	**RMSE^{s}**	$\bar{s}$	$\mathcal{E}(\bar{s})$	$\hat{\kappa}$	$\mathcal{E}(\hat{\kappa})$
0.1	0.661	0.925	0.314	0.049	0.158	0.831	0.225
0.2	0.666	0.916	0.308	0.061	0.147	0.800	0.194
0.3	0.654	0.968	0.299	0.081	0.127	0.750	0.144
0.4	0.662	0.943	0.297	0.085	0.122	0.731	0.125
0.5	0.662	0.896	0.296	0.088	0.120	0.720	0.114

Table 6: Adjusting the variability of MCDP models by dropout rate (dp).

I Further visualised examples: Emotion Class Labelling

This section shows additional visualized examples for emotion class labelling when human annotators reach a consensus (Figure 6 (a)(b)), diverge (Figure 6 (c)(d)), and give distinct labels (Figure 6 (e)). Aligned with the findings in Sec. 5.4, the proposed S-CNF can better simulate the aggregated behaviour as well as the variability of human annotations in all cases.

J Further visualised examples: Toxic Speech Detection

This section shows visualized examples for toxic speech detection when all three human annotators provide the same label (Figure 7 (a)(b)), one of them gives a different label (Figure 7 (c)(d)), and all three annoators give distinct labels (Figure 7 (e)). Similar to the observations of the emotion class labelling task, the proposed S-CNF can better simulate the aggregated behaviour as well as the variability of human annotations in all cases.

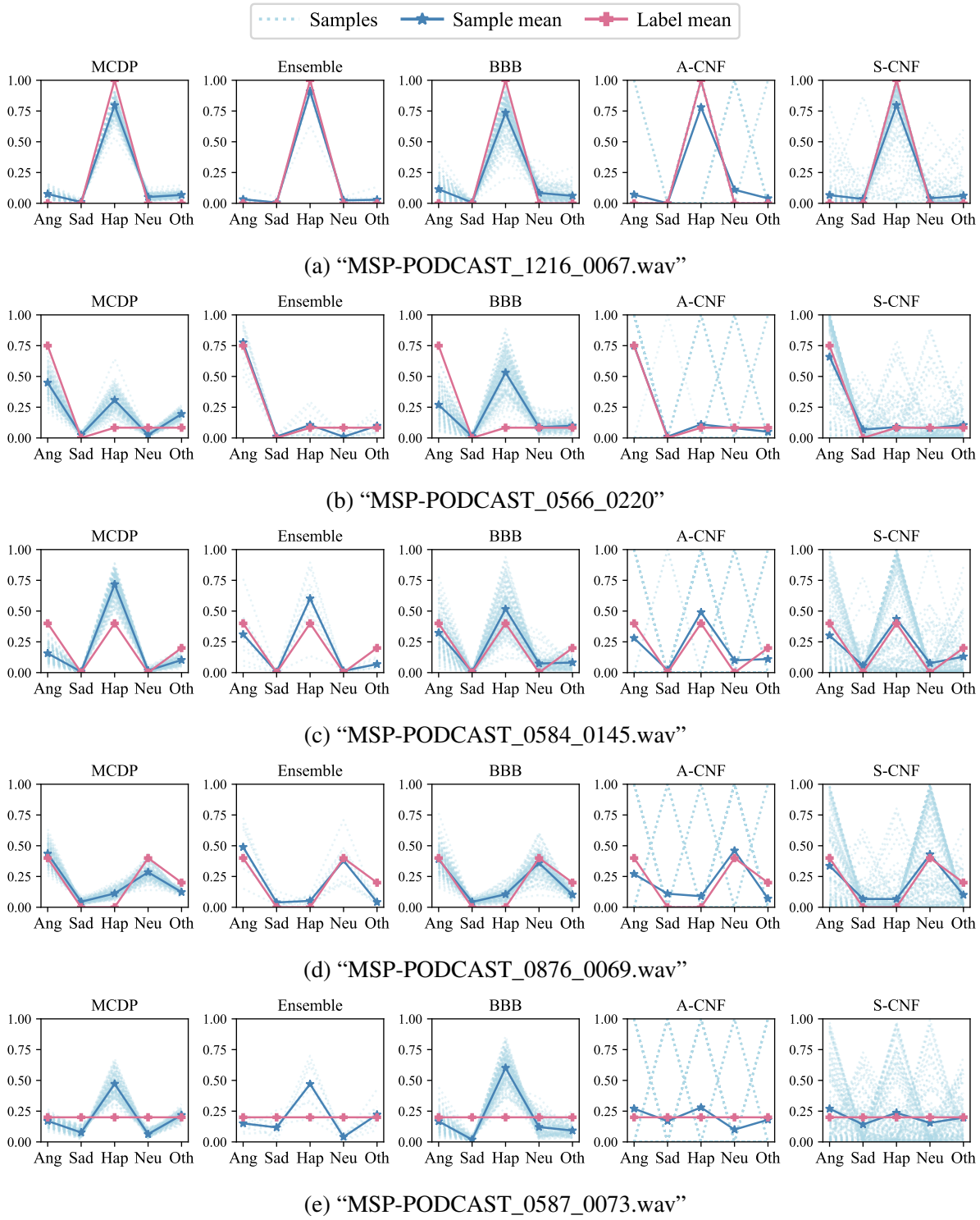

Figure 6: Additional visualized examples for emotion class labelling. The y-axis corresponds to the probability mass. Each sample is a categorical distribution. The probability mass values of different categories in each categorical distribution are connected for the purpose of better visualization.

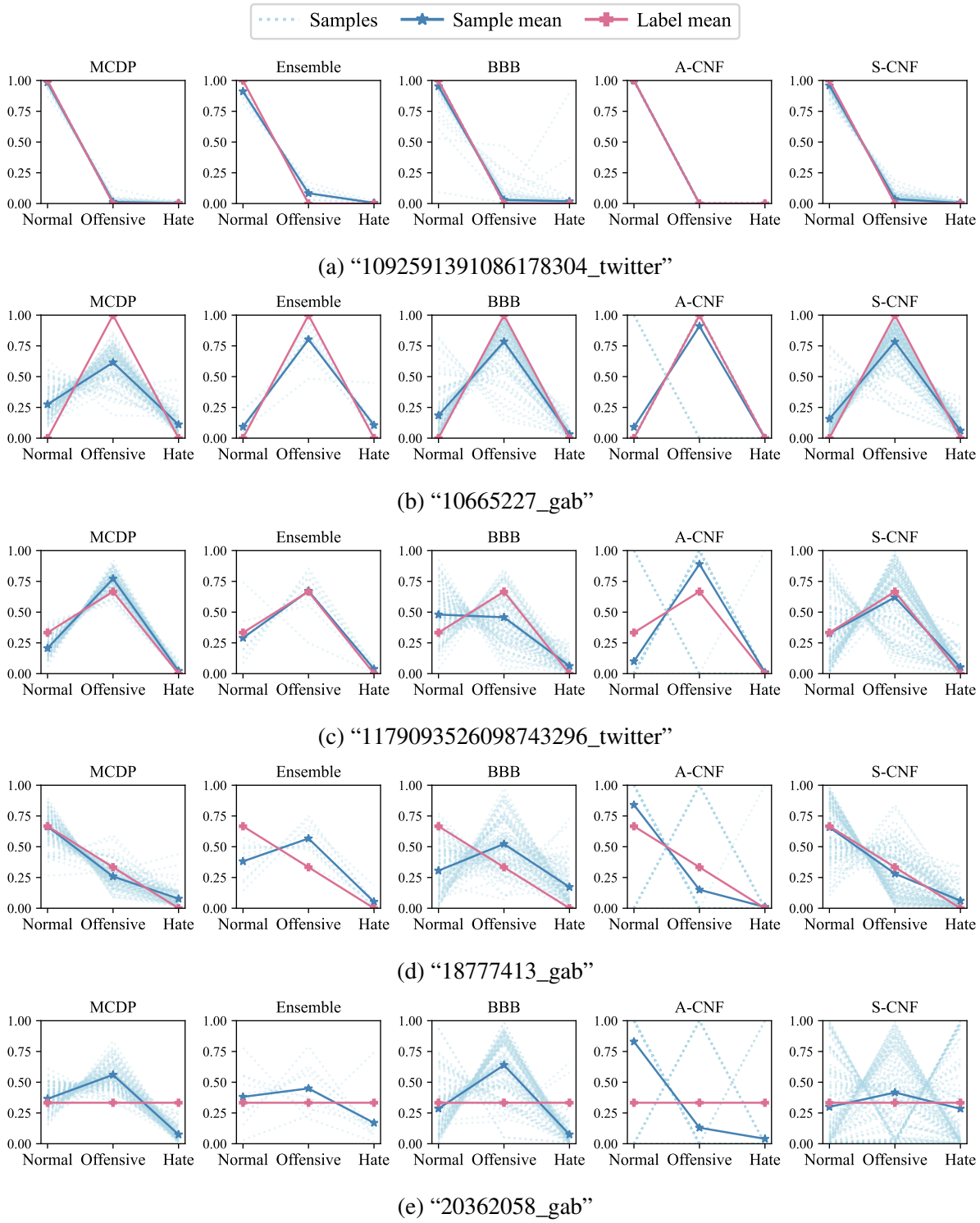

Figure 7: Additional visualized examples for toxic speech detection. The y-axis corresponds to the probability mass. Each sample is a categorical distribution. The probability mass values of different categories in each categorical distribution are connected for the purpose of better visualization.