

SeedBERT: Recovering Annotator Rating Distributions from an Aggregated Label

Aneesha Sampath,¹ Victoria Lin,¹ Louis-Philippe Morency¹

¹ Carnegie Mellon University
{aneeshas, vlin2, morency}@cs.cmu.edu

Abstract

Many machine learning tasks—particularly those in affective computing—are inherently subjective. When asked to classify facial expressions or to rate an individual’s attractiveness, humans may disagree with one another, and no single answer may be objectively correct. However, machine learning datasets commonly have just one “ground truth” label for each sample, so models trained on these labels may not perform well on tasks that are subjective in nature. Though allowing models to learn from the individual annotators’ ratings may help, most datasets do not provide annotator-specific labels for each sample. To address this issue, we propose SeedBERT, a method for recovering annotator rating distributions from a single label by inducing pre-trained models to attend to different portions of the input. Our human evaluations indicate that SeedBERT’s attention mechanism is consistent with human sources of annotator disagreement. Moreover, in our empirical evaluations using large language models, SeedBERT demonstrates substantial gains in performance on downstream subjective tasks compared both to standard deep learning models and to other models that account explicitly for annotator disagreement.

Introduction

Subjectivity uncertainty is uncertainty that arises from data with subjective or ambiguous labels—labels on which human annotators themselves can disagree. Many affective computing tasks, including automatic emotion recognition and hate-speech detection, are highly subjective due to their heavy reliance on human opinion. Different people may perceive a speaker’s toxicity level or emotional intent differently. If asked to rate a subjective sample, a group of five annotators may very well provide five different labels.

In contrast, machine learning datasets usually provide a single label for each sample. Labels for machine learning tasks are often crowd-sourced through survey platforms, where researchers request multiple annotators to rate each sample. Prior to a dataset’s public release, however, these individual annotators’ ratings are usually aggregated into one “gold standard” label. Flattening annotations into a single label results in loss of information about the subjectivity of the task, and consequently, models trained on aggregated labels

may not be able to achieve their desired level of performance on subjective tasks. This raises concerns for tasks such as emotion recognition, where perception is highly dependent on personal opinion (Zhang, Essl, and Mower Provost 2017; Gordon et al. 2021; Han et al. 2017).

To overcome such concerns and to better represent the nature of subjective tasks, machine learning models must be able to represent the subjectivity of a sample. In this paper, we propose models that can infer subjectivity as a distribution of possible human opinions. Due to the single-label nature of most standard machine learning datasets, we focus our efforts explicitly on models that are able to reconstruct these distributions *from single-label data*. In our modeling mechanism, we hypothesize that annotator subjectivity arises in significant part as a function of the portion of the sample on which the annotator decides to focus.

Illustrated Example. In the remainder of this paper, we take emotion recognition to be an example of a task with subjectivity uncertainty. Many existing emotion recognition datasets, including the language datasets CARER (Saravia et al. 2018) and Stanford Sentiment Treebank (SST) (Socher et al. 2013) provide only one emotion label.

We consider two sentences from the SST dataset with similar sentiment scores. In our first example, “You walk out of The Good Girl with mixed emotions – disapproval of Justine combined with a tinge of understanding for her actions,” the sentiment rating is 0.61, or slightly positive. The speaker expresses both negative and positive attitudes towards the film—“disapproval of Justine” and “a tinge of understanding for her actions,” respectively—and even explicitly notes mixed emotions in the sentence. These contrasting sentiments, however, are aggregated into just one score that leans positive (where the individual annotators might have considered the sentence to be very negative, slightly negative, very positive, etc.). On the other hand, in another sample from the dataset, “A smart, steamy mix of road movie, coming-of-age story and political satire,” the speaker expresses only positive emotions (indicating that the individual annotators all considered the sentence to be positive), but the sentiment score is 0.67, which is similar to that of the first example.

These score similarities demonstrate how the use of aggregated labels results in the loss of a nuanced representation of the sentiment. Models trained only on aggregated scores

may not learn a sentiment representation as well as those that are trained on the full distribution of annotator ratings.

Contributions. This paper’s contributions are threefold:

- We propose a deep language model (**SeedBERT**) that is capable of recovering original annotator distributions from a *single ground-truth label*. This model uses the recovered distribution to predict a label for the task.
- We introduce a new set of annotations (**MOSI-Subjectivity**) of human subjectivity and the mechanisms behind annotator disagreement. These annotations extend the CMU-MOSI sentiment analysis dataset (Zadeh et al. 2016).
- We present a detailed evaluation with both empirical performance analysis and a study of human agreement.

Related Work

Inter-Annotator Disagreement. Annotator disagreement can arise when tasks are inherently subjective. Affective semantics are particularly difficult to model since people’s biases, experiences, and knowledge lead them to form different opinions about the same data. This leads to performance degradation on tasks involving expressive language (Alm 2011). In subjective contexts, it can therefore be useful to move away from a single ground-truth label and to instead acknowledge that multiple answers may be valid.

Several prior works attempt to address this unmet need by explicitly modeling individual annotators. Models such as Jury Learning (Gordon et al. 2022) and HuBi-Medium (Kocoń et al. 2021) propose joint training of text embeddings and annotator embeddings to allow for predictions on the individual annotator-level. Although these approaches move toward personalized annotator modeling, they require not only that multiple annotators’ labels are available for each data sample, but also that annotator-identification information is available. This requirement contrasts with the typical crowd-sourced machine learning datasets, in which annotator-level information is not provided.

Modeling Perception Uncertainty. An additional body of work attempts to account for subjectivity by modeling perception uncertainty directly. Han et al. (2017) propose a multi-task framework that simultaneously predicts an emotion label and an estimate of annotator disagreement. The resulting model outperforms single-task emotion recognition models. Davani, Díaz, and Prabhakaran (2022) also propose a multi-task framework—one in which each task consists of modeling an individual annotator. The variance of the resulting annotator distribution is taken as the inter-annotator disagreement measure. However, both of these methods require access to annotator-level labels, which are not commonly available in public datasets, at training time. Moreover, the method proposed by Davani, Díaz, and Prabhakaran (2022) requires that each annotator label a large portion of the data.

MOSI-Subjectivity Dataset

To gain a better understanding of subjectivity uncertainty and the mechanisms by which it arises, we collected extensive human ratings of sentiment and subjectivity on an

emotion recognition task. We used these ratings to evaluate the empirical performance and human agreement of our proposed method, SeedBERT.

Dataset. We used video clips¹ from the Multimodal Corpus of Sentiment Intensity dataset (CMU-MOSI), a collection of over 2000 YouTube movie reviews with opinion-level annotations for sentiment and subjectivity.

Data Collection. We collected new labels for 500 randomly selected audiovisual CMU-MOSI samples using the *Prolific.co*² crowdsourcing platform. For each sample, we asked five high-quality annotators (approval rating $\geq 98\%$) to answer the following questions.

1. Is the speaker expressing *any* emotion in the video clip?
2. Select the emotion *most present* in the video.
3. Is the person expressing multiple emotions in the video? If so, select *all of the emotions present* in the video.

We intended the first question—which has relatively little ambiguity—to be a calibration question in order to establish the expected level of inter-annotator agreement in the absence of any subjectivity. This question was answered using a binary rating scale, with options *Yes* and *No*.

The second question, which requests that annotators mark the most salient emotion in the video clip, is one commonly asked in emotion recognition data collection procedures. CMU-MOSEI (Bagher Zadeh et al. 2018), for instance, used a similar question to collect its labels. As with CMU-MOSEI, the six Ekman emotions were presented as label options in a multiple-choice format.

The third question is a multiple-select question in which annotators were instructed to select all of the Ekman emotions that the speaker expresses in the video clip. This question allows raters to express greater nuance on the emotion recognition task rather than limiting them to a single response, which we felt could have implications for the subjectivity of the task.

After collecting a portion of the annotations and examining them, we hypothesized that disagreement might arise due to annotators attending to different parts of the sentence, where each part may emphasize a different emotion. To verify this hypothesis, we collected additional annotations for another 60 samples, wherein we asked raters to explain their selections. Again recruiting five high-quality annotators (approval rating $\geq 98\%$) to rate each sample, we asked:

4. On the previous page, you selected *X* as the emotion most present in the clip. Which words or phrases led you to your answer? Please highlight the portions of the transcript that led you to make your selection.

Analysis of Annotator (Dis)agreement

Quantitative Analysis. To determine the level of agreement among annotators for each of the questions asked, we

¹We initially collected data using transcripts only, but obtained poor inter-rater reliability on all questions.

²<https://www.prolific.co/>

calculated the Cohen’s kappa coefficient (Cohen 1960) using the *agreement* software package.³ To account for the multiple-selection setting of the third question, we treated the responses to each emotion as answers to a single binary question, then averaged Cohen’s kappa among all emotions.

Question	Kappa
Is the speaker expressing <i>any</i> emotion in the video?	0.744
Select the emotion <i>most present</i> in the video.	0.410
Select <i>all of the emotions present</i> in the video.	0.710

Table 1: Cohen’s kappa coefficients for unordered categories for the three survey questions.

We observe that annotators exhibit similar levels of agreement for the calibration question and the multi-label question (Table 1). However, agreement is substantially lower for the second question concerning the emotion most present in the sample. These results seem to indicate that annotators can acknowledge and agree upon the presence of other opinions, but greater subjectivity and disagreement arise when annotators are asked to select the most salient opinion.

Qualitative Analysis. Observing this discrepancy, we sought to investigate why annotators have high agreement when selecting all emotions present but low agreement when selecting the primary emotion. In Table 2, we provide selected examples to illustrate our analysis.

Sample	Main Emotions
these action sequences are certainly <i>a lot better</i> than the incoherent impossible to follow slop that we see in a typical michael bay movie	<i>Happiness</i> , Anger , Disgust
and i honestly think if we hadn't have seen watchman this would be the <i>best credit sequence</i> of the year	<i>Happiness</i> , Sadness , <i>Surprise</i>
but there are you know there're critics of movies that are sometimes too thoughtful in a way and <i>i think that would never let me go</i>	Disgust , <i>Surprise</i>

Table 2: Selected examples from the first survey. The first column is the transcript of the video clip shown to the annotators, and the second column is a set of the emotions selected as most present by the five annotators.

Given the transcript of the sample and the set of annotators’ responses to the emotion most present in the sample, we observe that different words or phrases within the sample may elicit different emotions. Taking the first example from Table 2, we can see that the italicized part of the sentence is characterized by a *happy tone*, whereas the bolded part has a **disgusted or angry tone**. While all annotators indicated that *happiness* and *disgust* or *anger* were present when answering the third question, they disagreed as to which of the

three was the primary emotion when answering the second question. This leads us to hypothesize that **annotators attend to different parts of the sentence**, which leads to low agreement when selecting the most salient emotion but high agreement when selecting all of the emotions present.

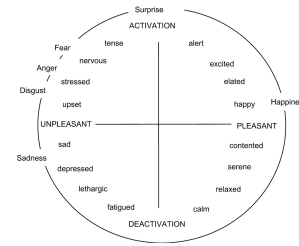


Figure 1: Circumplex model of affect.

As mentioned in the Data Collection section, we collected further annotations to verify this hypothesis, this time asking annotators to explain their selections for the most salient emotion by highlighting specific words and phrases in the transcript that led them to their decision. We found that our hypothesis holds when the emotions present are adjacent on the circumplex model of affect (Figure 1), in which similar emotions are adjacent to one another and opposite emotions are across from one another (Russell 1980).

Emotion	Supporting Evidence	Annotators
Happiness	a lot better	2
Anger	incoherent impossible to follow slop	2
Disgust	incoherent impossible to follow slop	1

Table 3: Selected example from the second set of surveys. Two annotators selected *Happiness* as the main emotion present, two selected *Anger*, and one selected *Disgust*.

Still using the first sample from Table 2 as an illustrative case, we observe in Table 3 that for the non-adjacent emotion pairs (*Happiness*, *Anger*) and (*Happiness*, *Disgust*), annotators cited different portions of the transcript to support their answer for the main emotion present, suggesting that they attend to different parts of the sentence. However, this finding did not hold for adjacent emotions, as the same portion of the sentence is provided as justification for both the *Anger* and *Disgust* labels. **This result may suggest that subjectivity may arise both from the part of the input attended to and from personal interpretation of the same portion of the input.** For the remainder of this work, we will focus on subjectivity due to the part of the utterance attended to.

Methods

SeedBERT. Our proposed method, SeedBERT (Figure 2), is premised on the idea that a collection of randomly initialized layer in a neural network will (by virtue of their different starting points) learn to attend to different parts of an input, even when trained on the same data. When the input is

³<https://github.com/jmgirard/agreement>

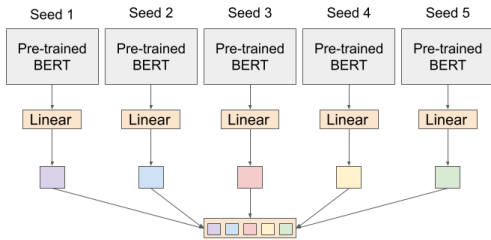


Figure 2: SeedBERT architecture. Pre-trained BERT encoders are paired with a linear layer initialized from a random seed.

inherently subjective, the linear layer—in attending to a particular portion—will represent the opinions and biases of an arbitrary annotator. Following this reasoning, SeedBERT is a collection of n identical networks consisting of pre-trained BERT encoders with a task-specific linear layer. Each network is trained on the same dataset with identical loss functions, with the only difference being that each network has a different random initialization of the final linear layer. Each network is meant to reflect one annotator, and SeedBERT’s final prediction is the majority vote of the networks.

As an ensemble of networks, SeedBERT bears certain methodological similarities to classical ensemble methods like bagging (Breiman 1996). Rather than using a collection of weak learners to improve predictive performance, however, SeedBERT leverages a powerful existing model—pre-trained BERT—and uses its “ensembling” to induce individual learners to attend to different portions of the input.

Evaluation. We compare our approach, SeedBERT, against four baseline models, two of which we describe previously in the Related Work section. These latter two models are trained on individual annotator ratings, which are not typically available in machine learning datasets. Due to their use of this additional training data, we can consider them to be oracle models that provide upper bounds on the ability of a model to learn annotator distributions. All models use pre-trained BERT as an encoder (Kenton and Toutanova 2019), with a final linear layer fine-tuned for the task via the Adam optimizer. We use a learning rate of 5×10^{-6} and fine-tune for 3 epochs. We evaluated our models on metrics accuracy and F_1 -score using 5-fold cross-validation and repeated this process 5 times with different splits of the data.

Baseline Models. Our first baseline model (**BERT**) reflects a system where there is no explicit attempt to account for subjectivity. This model consists of a single pre-trained BERT encoder with a task-specific final linear layer.

Our second baseline model incorporates Bayesian neural networks (**BNN**), which have previously been used to estimate prediction uncertainty due to their use of probabilistic rather than deterministic weights. Recent work has applied BNNs to speech emotion recognition in order to directly estimate annotator distributions (Raj Prabhu et al. 2022). We leverage a similar architecture of an encoder followed by a BNN, but use a transformer-based BERT encoder.

We also compare SeedBERT against two oracle models

trained on individual annotator ratings, which are not typically available in machine learning datasets. We use a Label Distribution Learning (**LDL**) framework (Geng 2016) and the **Multi-Task** framework proposed by Han et al. (2017) to determine the upper bounds of achievable model performance when modeling annotator distributions.

Results & Discussion

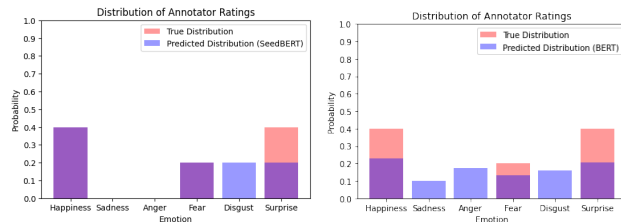


Figure 3: Example of SeedBERT predicted annotator distribution of emotions (left) and BERT predicted distribution (right). The predicted distributions are in blue, while the true annotator distributions are in red.

In our empirical evaluation, we observe in Table 4 that SeedBERT model outperforms all other models on the emotion recognition task. Notably, given only the aggregated label at training time, SeedBERT outperforms not only the baselines (standard pre-trained BERT and BNN) but also the two models trained on individual annotator data. These results suggest that SeedBERT’s mechanism of reconstructing annotator distributions is effective in improving performance on subjective tasks.

Through qualitative analyses, we find that an individual BERT model tends to predict uniform distributions of annotator ratings (i.e., uniform across emotions) irrespective of the input sample (Figure 3), possibly because it cannot fit well to low-agreement data. As seen in Table 4, this shortcoming has clear negative implications for its performance on the emotion recognition task. On the other hand, the SeedBERT approach moves toward reflecting annotators’ biases, as each random initialization can reflect the preferences of an arbitrary annotator. As shown in Figure 3, the annotator distribution predicted by SeedBERT much more closely reflects the true annotator distribution.

We further find that certain random initializations of the individual SeedBERT networks tended to bias toward a subset of the Ekman emotions—much in the same way that a human annotator might have a tendency to favor certain emotions in their ratings. These observations align with our hypothesis that the random initializations may serve as proxies for individual annotator perceptions or biases.

Conclusions

In this paper, we analyze human subjectivity and the mechanisms by which annotator disagreement arises. We propose SeedBERT, a novel method capable of recovering original annotator distributions from a single ground-truth label to ultimately improve performance on downstream tasks. Via

Model	Accuracy	F_1 -Score
BERT	.204	.102
LDL	.284	.0821
Multi-Task	.180	.0954
BNN	.150	.132
SeedBERT	.308	.183

Table 4: SeedBERT outperforms all baselines when averaged across 5 iterations of 5-fold cross validation.

SeedBERT, we find that manipulating the random initializations of pre-trained models can serve as an effective proxy for modeling varying annotator perceptions.

Our findings should be considered in light of several points. First, our extension to CMU-MOSI consisted of a relatively small number of samples (500) for fine-tuning a large language model. Second, fewer samples were rated for *anger* and *fear*, which could limit the efficacy of our method when evaluating these emotions. Third, although CMU-MOSI was designed to be a multimodal dataset, we implemented SeedBERT and its comparison methods as unimodal language models, which likely limits their predictive ability. Fourth, the validation set accuracies and F_1 -scores are relatively low across all models, which warrants further exploration. Finally, though random initializations can effectively represent a collection of arbitrary annotators, there are cases where prior information about the annotators’ preferences is known (for example, an annotator’s favorite movie genre). In these cases, SeedBERT may be improved by incorporating this prior information to represent a *specific* annotator distribution rather than an arbitrary one.

References

- Alm, C. O. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 107–112.
- Bagher Zadeh, A.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2236–2246. Melbourne, Australia: Association for Computational Linguistics.
- Breiman, L. 1996. Bagging predictors. *Machine learning*, 24(2): 123–140.
- Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.
- Davani, A. M.; Díaz, M.; and Prabhakaran, V. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10: 92–110.
- Geng, X. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7): 1734–1748.
- Gordon, M. L.; Lam, M. S.; Park, J. S.; Patel, K.; Hancock, J.; Hashimoto, T.; and Bernstein, M. S. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*, 1–19.
- Gordon, M. L.; Zhou, K.; Patel, K.; Hashimoto, T.; and Bernstein, M. S. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21. New York, NY, USA: Association for Computing Machinery.
- Han, J.; Zhang, Z.; Schmitt, M.; Pantic, M.; and Schuller, B. 2017. From Hard to Soft: Towards More Human-like Emotion Recognition by Modelling the Perception Uncertainty. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM ’17, 890–897. New York, NY, USA: Association for Computing Machinery.
- Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.
- Kocoń, J.; Gruza, M.; Bielaniec, J.; Grimling, D.; Kancierz, K.; Miłkowski, P.; and Kazienko, P. 2021. Learning personal human biases and representations for subjective tasks in natural language processing. In *2021 IEEE International Conference on Data Mining (ICDM)*, 1168–1173. IEEE.
- Raj Prabhu, N.; Carbajal, G.; Lehmann-Willenbrock, N.; and Gerkmann, T. 2022. End-To-End Label Uncertainty Modeling for Speech-based Arousal Recognition Using Bayesian Neural Networks. 151–155.
- Russell, J. A. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6): 1161.
- Saravia, E.; Liu, H.-C. T.; Huang, Y.-H.; Wu, J.; and Chen, Y.-S. 2018. CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3687–3697. Brussels, Belgium: Association for Computational Linguistics.
- Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A.; and Potts, C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. Seattle, Washington, USA: Association for Computational Linguistics.
- Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6).
- Zhang, B.; Essl, G.; and Mower Provost, E. 2017. Predicting the Distribution of Emotion Perception: Capturing Inter-Rater Variability. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI ’17, 51–59. New York, NY, USA: Association for Computing Machinery.