# Impact of Annotator Demographics on Sentiment Dataset Labeling

YI DING, Georgia State University, USA
JACOB YOU, Westlake High School, USA
TONJA-KATRIN MACHULLA, TU Dortmund, Germany
JENNIFER JACOBS, University of California Santa Barbara, USA
PRADEEP SEN, University of California Santa Barbara, USA
TOBIAS HÖLLERER, University of California Santa Barbara, USA

As machine learning methods become more powerful and capture more nuances of human behavior, biases in the dataset can shape what the model learns and is evaluated on. This paper explores and attempts to quantify the uncertainties and biases due to *annotator* demographics when creating sentiment analysis datasets. We ask >1000 crowdworkers to provide their demographic information and annotations for multimodal sentiment data and its component modalities. We show that demographic differences among annotators impute a significant effect on their ratings, and that these effects also occur in each component modality. We compare predictions of different state-of-the-art multimodal machine learning algorithms against annotations provided by different demographic groups, and find that changing annotator demographics can cause >4.5% in accuracy difference when determining positive versus negative sentiment. Our findings underscore the importance of accounting for crowdworker attributes, such as demographics, when building datasets, evaluating algorithms, and interpreting results for sentiment analysis.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; User studies; • **Computing methodologies → Machine learning**.

Additional Key Words and Phrases: demographics; annotator; annotation; sentiment; multimodal; machine learning; artificial intelligence; bias; emotion; fairness; dataset; crowdworker.

**ACM Reference Format:**
Yi Ding, Jacob You, Tonja-Katrin Machulla, Jennifer Jacobs, Pradeep Sen, and Tobias Höllerer. 2022. Impact of Annotator Demographics on Sentiment Dataset Labeling. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 519 (November 2022), 22 pages. https://doi.org/10.1145/3555632

## 1 INTRODUCTION

Multimodal sentiment analysis presents the challenge of computationally determining how humans would emotionally interpret a given input. For example, what is the sentiment expressed by a speaker in a video? This problem is critical for building richer human-computer interaction experiences and providing automated assistance to people, for example. Leveraging the power of deep learning, researchers have made progress modeling sentiment in any modality of input including text, video, audio, or some combination of multiple modalities. To train complex, non-linear deep learning

Authors' addresses: Yi Ding, yiding@gsu.edu, Georgia State University, Atlanta, USA; Jacob You, jyoucs@gmail.com, Westlake High School, Westlake Village, USA; Tonja-Katrin Machulla, tonja.machulla@tu-dortmund.de, TU Dortmund, Dortmund, Germany; Jennifer Jacobs, jmjacobs@mat.ucsb.edu, University of California Santa Barbara, Santa Barbara, USA; Pradeep Sen, psen@ece.ucsb.edu, University of California Santa Barbara, Santa Barbara, USA; Tobias Höllerer, holl@cs.ucsb.edu, University of California Santa Barbara, Santa Barbara, USA.




2573-0142/2022/11-ART519
https://doi.org/10.1145/3555632

models, researchers have created datasets consisting of thousands of video examples labeled according to sentiment [8, 46, 49, 81–83]. Most of these datasets consist of example videos of "talking head" speakers. Each example is labeled independently by a set of human annotators who are asked to gauge the emotion of the speaker (whether they are saying something positive, negative, or neutral, for instance). As research in sentiment analysis has progressed, AI models that classify sentiment have been applied to a range of decision making pipelines and applications, including emotional chat support [69] and determining hate speech [78]. Furthermore, as these technologies advance and become more pervasive, it can drastically alter how we live our daily lives from seeking out medical help [28] to and changing our work environment [75].

To create the datasets to train the models in these systems, researchers often rely on hiring crowd workers through services such as Amazon's Mechanical Turk to label the data. There has been extensive work examining how to leverage crowd workers to obtain quality labels at scale [43, 65], and sentiment is frequently labeled in the same manner. However, sentiment is highly subjective, and when appraising sentiment of others, psychologists have found that our experiences and opinions play an important role [1, 7, 14, 23, 37, 42, 45, 50, 56, 62]. Since these studies have shown that demographics can be used to capture differences in sentiment appraisal, it therefore suggests that differences in annotator demographics can lead to differences in their interpretation of sentiment.

To attempt to control for the impacts of demographics, researchers have developed previous datasets that balance some variables. For example, balancing for a 50-50 distribution of male and female speakers [83]. Newer datasets also occasionally provide demographic information of the speakers or subjects in the dataset [81, 82] as a means to aid models in making more informed decisions. Other works have also examined the biases inherent to the contents [36, 76, 79]. However, few works have examined biases due to annotators' demographic backgrounds [2, 32, 53]. To the best of our knowledge, no works have examined annotator biases for sentiment from a multimodal perspective.

If annotator demographics impact sentiment, then any results gleaned from a dataset that does not control for annotator demographics at the time of creation will be biased and skewed in addition to all the other biases that such datasets already exhibit [11, 30, 77] by any "unbalanced" (defined relative to specific application needs and goals) distribution of annotators. Therefore, if the demographics of the annotators did not match the distribution of, say, the general population, then results and analysis using that dataset might not be applicable to the general population. Furthermore, models and evaluations using these datasets would reflect the opinions of those who perform crowdwork versus those who do not. For systems which make decisions based on these models, this would mean lower efficacy for certain groups of users. However, as these technologies become increasingly critical in commonplace technological systems, it would not be far-fetched to notice a disenfranchisement of specific demographics of people [5]

Understanding the potential role of annotator demographic is critical in informing decisions about how we use and trust sentiment analysis technologies going forward. We attempt to provide some answers to this and quantify the impact of demographics on sentiment analysis datasets. We re-labeled the well-established MOSEI dataset [83]: a dataset of "talking head" speakers scraped from YouTube. Using a crowd-sourced labeling process that took the annotators' demographic information into account, we produce a rich annotation that includes 5 times more annotators per video than the original MOSEI dataset. We also collected detailed demographic information for all annotators in our relabeled dataset. Using this dataset, we conduct statistical experiments to establish and begin to quantify the impact of demographic background on sentiment labeling. From this analysis we found that annotator sentiment varies (with statistical significance) based on demographic factors such as age, gender, ethnicity, and educational level. Our results suggests

that decisions derived from AI should be used cautiously and a strong need for interdisciplinary collaboration for more inclusive AI development.

Our work provides the following contributions:

- We present a large set of annotations for multimodal sentiment analysis containing rich demographic information. Additionally, we provide annotations for independent modalities (text, audio, visual) in addition to their combined annotations.
- We show that demographics have a significant effect on sentiment labeling, and show that this effect generalizes across all component modalities: text, visual, audio, and their combination. We find noticeable differences in label agreement and ratings for different modalities. These results suggest that decisions derived from AI results should be used cautiously. In particular, they should consider the parameters for data collection that may introduce unintended biases.
- We show that algorithmic claims of sentiment classifier improvement can vary greatly due to demographics, observing up to 4.9% change in absolute algorithmic performance when sampling for various sub-populations. This exceeds the improvement claims over state-of-the-art of most recent sentiment classification machine learning papers. We additionally show the ability for our gathered labels to be used as an improved evaluation metric to account for demographic biases. Data is released for public use.

## 2 RELATED WORK

We examine relevant literature to motivate our work. We first examine modern advancements in multimodal machine learning that is applied to sentiment or emotion classification. We discuss datasets these models are trained with and their annotation process. We then examine how demographics can influence the emotion appraisal process. Finally, we examine ways in which works have attempted to quantify and mitigate the demographic of annotators.

### 2.1 Multimodal Machine Learning

Enabling machine learning for multimodal data has been explored in many domains over a long period of time [4]. Many machine learning techniques have been applied on the task of sentiment classification [27, 70]. As transformer-based architectures have become very popular recently, some recent techniques have also explored their use in multimodal settings. Originally proposed in [68] for neural machine translation (NMT) tasks, they have demonstrated superior performance on multiple benchmark problems such as in image classification [21] and action recognition [47]. The basic functionality is to apply layers of self-attention, on sequential representations. Recent attempts by researchers to enable multimodal modeling on transformers via cross-modal attention have been successful for sentiment analysis [20, 66]. Inspired by work which showed that shifting one modality (language) using representations from other modalities improves performance, [70], MAG-XLNet [55] incorporates the ability for fine-turning on multimodal data on a transformer-like model built on top of XLNet [80]. XLNet is an extension of transformer based methods that enables learning over longer sequences and the ability to better model the context dependencies.

### 2.2 Multimodal Machine Learning Datasets

There is a long line of work for building large scale datasets for machine learning. There have been numerous works on the development of large scale datasets for the vision, language, and multimodal domains. Many datasets have been gathered over the years to explore sentiment or valence: text based, visual, audio, and via their multimodal combination [81–83]. Additional datasets using modalities such as pose [8, 49] and EEG [19, 38] have also been created and analyzed. Datasets built around continuous representations have also been explored [39].

| | # samples | Mean annotators per sample | Total annotations | Scale | Annotator demographics? | Per-modality labels? |
|---|---|---|---|---|---|---|
| **Ours** | 500 | 15 | 30,000 | 7-pt likert | Yes | Yes |
| **CH-SIMS** [81] | 2281 | 3 | 27,372 | 7-pt likert | No | Yes |
| **MOSEI** [83] | 23,453 | 3 | 70,359 | 7-pt likert | No | No |
| **SEWA** [39] | 1990 | 5 | Continous | Continuous | No | Yes |

Table 1. High level statistics of recent datasets for sentiment or emotion analysis. Our annotation effort produced more ratings per sample and also contains detailed annotator information. We also provide per-modality labels and have a comparable number of total annotations in the entire dataset. This type of annotation enables us to perform in-depth analysis on demographic effects and is comparable to large-scale machine learning datasets by annotation count.

Due to their large size, these datasets are typically labelled via crowdsourcing platforms such as Amazon Mechanical Turk. One frequently used method for obtaining quality labels is the use of multiple annotations and taking the mean or majority class. While many existing datasets simplify annotations to a single ground truth emotion or sentiment label, likely due to a lack of annotators per sample, emotion representation is not necessarily discrete. Representations such as [52] describe emotions in a continuous space. In this work we compare the mean label for ease of comparison with prior art, however, the scope of our data collection enables us to represent labels as a distribution (with mean and variance).

### 2.3 Demographic Effect on Emotional Appraisal

The studies of how emotions are interpreted have a long history in psychology. Demographics such as gender [23, 50, 64], age [1, 45], culture [7], economic background [54], etc., play a particularly large role. There are significant differences in the emotion expression and appraisal as a result of these factors. Combinations between multiple demographic variables have also been considered, such as in age and culture [1]. For example, Plant et al. [50] showed that people typically rate women sadder than men, and that they demonstrated a wider variety of emotions. Fischer et al. [23] found that women's experiences of emotions were modulated by cultural background. Many works have also explored gender stereotypes beyond this [7, 34].

Cultural backgrounds have also played a role. Brody [7] presented data showing that emotion expressiveness across cultures are different. Davis et al. [14] demonstrated that elicited emotional responses are different between participants of Chinese versus American culture between men and women. Age has also been well studied: Mitchell et al. [45] found that older adults are less accurate at interpreting prosodic emotion cues, and follows numerous previous works studying the age-related decline for identifying emotional cues [58]. Additional differences in age demographics between rater and poser were also discussed by Riediger et al. [56], and that emotional expression by older posers were more difficult to read.

### 2.4 Annotator Bias

The study of annotator demographics and its relationship to machine learning dataset creation is not new. Many techniques have tackled this during data acquisition [24, 25] as well as during model development [33, 48]. Works such as Wauthier and Jordan [71] proposed a framework to mitigate worker biases and downstream effects on model performance. Asking workers to think about other workers responses as demonstrated by Shaw et al. encouraged workers to provide more objective

annotations [63]. This inspired Hube et al. [32] to develop a method for intervention to overcome the strong influence that personal opinions have during annotation of subjective datasets. Chung et al. [10] recently conducted a systematic evaluation of different approaches for obtaining ground truth labels. Most recently, techniques have been proposed by Chen et al. [9] have attempted to capture annotation uncertainty as well as improve consistency by improving the annotation task design.

Furthermore, recent efforts to combat bias have been a topic of focus in the natural language processing domain. In particular, it has been observed that for datasets pertaining to hate-speech, the demographics of annotators play a large role [2, 26, 35, 40, 60, 72]. These results are echoed by Prabhakaran et al. [53] who found that annotators for hate-speech [35], sentiment [17], and emotions [15] for the language modality contained bias due to annotator demographics. That is, aggregated labels did not properly capture the perspectives of annotators from varying demographic groups. In fact, the impact of annotator demographics have also been observed when obtaining credibility ratings for news [6]. Some recent works have begun to quantify and mitigate this effect. Gordon et al. [29] introduced a metric to correct metrics based on the assumption that annotators will provide inaccurate answers with some chance.

We build upon these works by providing more detailed data as well as analysis for the multimodal domain. Additionally, we examine the scale of these effects empirically using state of the art techniques.

## 3 EXPERIMENTAL DESIGN

We examine the impact of how demographics of annotators impact label distribution. In other words, do the demographics of annotators matter when labeling sentiment data? Our hypothesis is that by grouping annotations based on the demographics of annotators who provided them, we can drastically alter the "ground truth" label distribution, to the point that these differences might even outweigh differences among competitive machine learning classifiers attempting to approximate this ground truth. Furthermore, with the intuition that model performance and evaluation will be strongly affected, we examine how annotations from strategically selected demographic subgroups can be used to create a demographic bound on performance. Naturally, this requires gathering data that can capture nuances of demographic effects and is also capable of being used by recent machine learning architectures for evaluation.

We divide the experiment into two parts: 1) We first conduct a large scale annotation ($>$ 1000 annotators) of videos used for multimodal sentiment classification, and 2) We additionally conduct a set of statistical experiments to determine the significance and impact of annotator demographics on dataset labels.

We choose to evaluate *sentiment*, in this case positive versus negative speaker stance (as evidenced by language, speaker video and speaker audio), due to the simplicity in the annotator decision making (one single axis) compared to more complex emotion measures. Agreement scores among raters are generally much lower for emotion datasets than for sentiment (positive vs negative) only [41]. We perform our investigations in the chosen domain to provide a strong baseline, as more obvious annotations should intuitively be least affected by demographic differences. Furthermore, as we wish to examine whether the effects of modality would modulate any annotator biases, we additionally gather annotations for the individual component modality for each sample. We present our process for data collection in Section 4.

We build our dataset by randomly sampling from the test set for MOSEI based on the split from [66] and [55] for re-annotation. This is a very suitable dataset for our purposes. The fact that we are extending an existing established dataset means that we have a baseline set of annotations to compare to when performing analysis of the new annotations. Another benefit is that the machine

learning research community actively produced and evaluated classifiers for MOSEI, which can be used in our evaluation [20, 55, 66, 70]. By changing the demographic distribution of the test set, we determine the approximate effect that this would have on machine learning models. We discuss and conduct thorough experiments in Section 5. In summary, our goals are to 1) provide a set of annotations large enough for machine learning evaluation and to understand demographic influences and 2) provide empirical evidence for the scale of annotator demographic effects on sentiment dataset labeling across modalities.

## 4 DATA COLLECTION

We now describe the data collection process for our large scale study to examine the influences of annotator demographics on unimodal and multimodal sentiment-dataset annotation. We first describe the video samples used for annotation in Section 4.1. We then discuss the platform used for recruiting participants (Section 4.2), and the collection interface (Section 4.3). Lastly, we discuss our way to improve the demographic distribution of the annotators in Section 4.4.

### 4.1 Multimodal video samples

We build our dataset using 500 videos segments randomly sampled from the Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) dataset [83]. It is one of the largest multimodal sentiment analysis datasets to date, and is highly regarded in the domain. The dataset is gender-balanced for male/female speakers. All sentences are annotated and randomly selected from various topics and monologues. The dataset contains over 23k video segments of 7.28 seconds long. Each segment was annotated by 3 annotators on a 7-point Likert scale. This resulted in over 70k total annotations.

To answer questions regarding modality effects, we split each video into its component modalities: audio, video, text, and their combination. This results in a total of 2000 different samples. Annotators are randomly assigned 30 of the 2000 samples for annotation. We ask more than 1000 annotators to provide ratings and results in approximately 15 annotations per sample. This enables us to to capture the per sample demographic and population effects on sentiment annotation for all modalities.

### 4.2 Prolific crowdsourcing

We use Prolific to gather annotations for the samples. Prolific is a crowdsourcing tool similar to Amazon's Mechanical Turk available to countries within the Organisation for Economic Co-operation and Development (OECD). In addition to being able to designate tasks to crowdworkers, Prolific gathers demographics of participants and makes this information available to researchers. Researchers can filter for a specific participants based on demographic background. We chose to go with Prolific as our crowd sourcing tool due to this accessibility of diverse participant information and its strict verification process (via government issued identification). We obtained crowdworker ethnicity data, country of birth, employment status, student status, gender identity, fluent languages, highest education level, immigration, whether participants were mono/multi culture, their nationality, and household income.

In addition to filtering for data we need for analysis, we filter participants by parameters to maintain data quality. In particular, we only accept workers with greater than 97 percent approval, who are fluent in English with no language related disorders. It is also required that participants can see video and hear audio.

This work involves reannotating part of an existing dataset consisting of non-offensive video footage of movie and other reviews. We do not collect demographic information ourselves or have access to any private information of the annotators. The participants have further agreed that some high level demographic information will be shared and used for research purposes when signing up

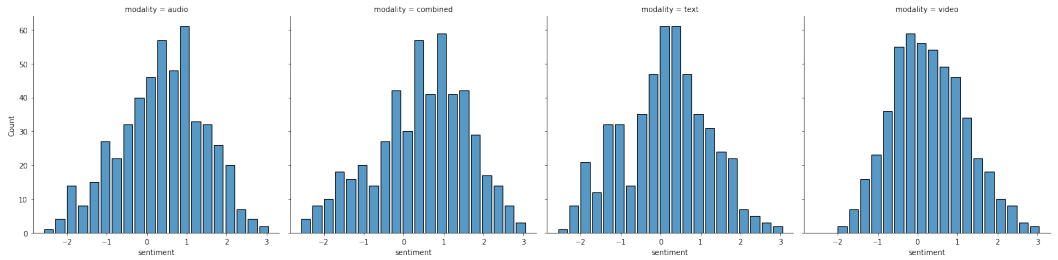



Fig. 1. Distribution of sentiment ratings by modality. The distributions are similar to the original ratings, however noticeable differences exist when examining different modalities.

for our study through Prolific. The annotation task is quick, and we experimentors did not interact directly with any annotators. As a result of these factors, our institution's IRB has determined our methodology to be exempt human subjects research. We recognize demographic properties of participants is sensitive information and follow protocols to protect the privacy of the annotators.

The representations of demographic properties are limited by the availability of information provided by Prolific and exclude certain populations from analysis. We encourage the reader to interpret results with these factors in mind.

### 4.3 Collection interface

We follow the same annotation process described in detail by Liang et al. [41] for data collection interface to reduce variability between the experiments. The participants are presented with a series of bullet points explaining the task of sentiment analysis, as we did not have access to the original training videos used by Zadeh et al [83]. We describe sentiment as the speaker's attitude towards the topic of their discussion. We also asked the annotators to rate the sentiment of the speaker, and not their own opinions. As sentiment labeling is a frequent task on crowd working sites, we expected most annotators to be able to accomplish the task with minimal training. We did not provide too much training as we, in accordance with previous annotation goals, wanted to avoid the over-training of subjects and wanted to maintain the in-the-wild goal of the dataset.

After receiving directions, the participants are given 30 random samples to annotate. We asked participants to provide ratings on a 7-point Likert scale for sentiment from highly negative (-3) to highly positive (3). We further ask the participant to rate the gender of the speaker as well as ask if any samples failed to display properly. We match the gender assessment of each sample with the original annotations as an additional way to maintain data quality. While we do not use this assessment of the speaker in our work directly for analysis, we anticipate future work to explore relationships between annotator demographics and data properties.

We estimated the time to complete 30 questions to be approximately 13 minutes or approximately 20 seconds per sample with some extra time for the directions. We targeted payment to be approximately $9.50 per hour as this is the good pay threshold set by Prolific. However, our actual hourly rate ended up being approximately $13 per hour as text was much faster to label and the average taken length for 30 questions ended up being 10 minutes instead. We did not reduce this pay as we did not wish affect the task adoption or completion rate [44].

### 4.4 Improving demographic diversity of annotations

We divide the annotation process into two phases during the summer of 2021. In the first phase, we asked 500 participants from the US to provide annotations without controlling for any specific

| Age | Before boosting | After boosting | Prolific dist. |
|---|---|---|---|
| **≤ 20** | 27% | 21% | 17% |
| **21-30** | 58% | 46% | 50% |
| **31-40** | 8% | 16% | 18% |
| **> 40** | 6% | 17% | 14% |

(a) Demographic distribution by age. By gathering more data from under-represented groups (> 30) we reduced the demographic skew. Older populations were less active in picking up our task and likely amplified the distribution bias.

| Gender | Before boosting | After boosting | Prolific dist. |
|---|---|---|---|
| **Female** | 79% | 73% | 73% |
| **Other** | 21% | 27% | 27% |
| **Ethnicity** | | | |
| **White** | 73% | 67% | 74% |
| **Other** | 27% | 33% | 26% |

(b) Demographic properties by gender and ethnicity before and after gathering data from under-represented groups.

Table 2. Summary of demographic properties of the annotations. We report the proportion of labels before and after applying a boosting process to increase the number of under-represented demographic groups of online annotators. Each annotator provided 30 labels and we obtained data from approximately 1000 people. We also provide the approximate active proportion of annotators available to researchers as reported by Prolific. These distributions are in agreement with previous findings that a majority of crowd workers are white, young (under 30), and female. More data is collected from under-represented groups by filtering for candidates in these groups, i.e., annotations over age 30, annotators who are not female, and annotators who are not white. The boosting process provided a noticeable benefit to the representation of the dataset. However, annotations by underrepresented groups typically took longer, and participants were less actively picking up our annotation task. This would explain some of the more dramatic demographic skews in age and gender before applying the boosting process.

demographic backgrounds. In the second phase, we gather data from an additional 500 annotators by restricting certain demographic backgrounds. The process for restricting annotations from certain demographics is the same as quality related properties such as annotator approval rates. Our task is only visible to particular populations that match a pre-specified group. That is, we ask for annotators who are older than 30, who are not female, and who are not white. We keep quality related filters such as approval rates, language skills and others the same. We perform this restricted group annotation process independently for each group. We report the demographic proportions from this process in Table 2. We see that the dataset is heavily skewed before this restriction process, and that the representation of smaller groups is boosted afterwards. We also report the approximate overall Prolific distribution of annotators. However, we found that underrepresented groups typically participated far less actively in the annotation process and thus amplified some dataset skews seen in age and gender.

For our study, participants who did not provide demographic information were removed from the list of pool of potential annotators. We further limited our scope of research to participants within the US to limit potential geographic effects. Including multiple geographies would also exacerbate the long-tailed distributions of demographic properties due to additional variables. However, since a large portion of the active participants on Prolific appear to live in the US, we still had a sizeable population for recruitment. Additionally, as a vast majority of the active users on Mechanical Turk are also from the US [18], we anticipated significant demographic overlap with the original labels. After applying these filters, there were approximately 50k participants who were active in the last 90 days prior to data collection.

In general we found that when boosting for specific demographics, it took longer before our annotations were completed. We also found that by only removing one demographic from consideration, such as only permitting annotators who were not female, that the other demographic properties (gender and ethnicity) were still heavily skewed. Although using more specific filters may help, such as filtering for non-female and non-white, we chose to use more general filters as some population demographics were very small that we did not want to introduce other population effects. We did not notice any differences in annotation quality for different demographics.

## 5 EXPERIMENTS AND RESULTS

We conduct experiments and present results to first provide a summarized view of the gathered data for context to interpret the results Section 5.1. We present the significance of demographic effects in Section 5.2. We then perform a series of experiments to explore the impact of these effects. We first examine the effects of this via Monte Carlo simulation experiments in Section 5.3. We then analyze how inter-annotator reliability is impacted in Section 5.4. Lastly, we analyze the impact that this has on state-of-the-art learning algorithms for this task in Section 5.5.

### 5.1 Data overview

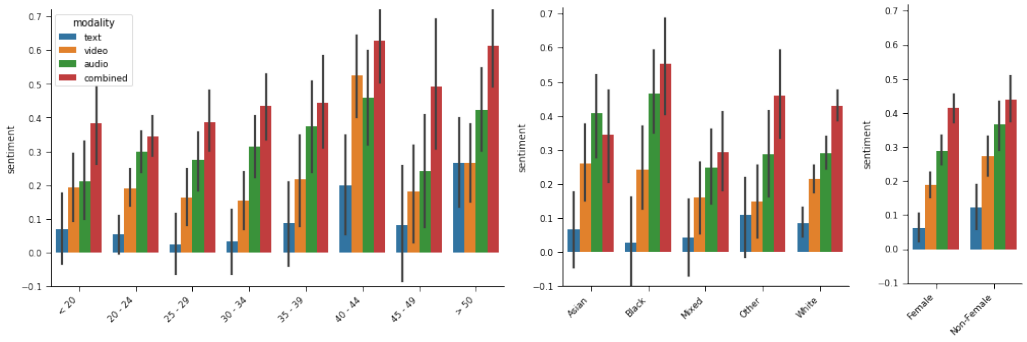
A total of 1034 annotators participated in our annotation task. We removed participants who did not complete the task by answering all 30 questions, completed the task too quickly, or experienced connectivity issues. A final 886 annotators completed the task with the distribution of population shown in Table 2.

We find that without controlling for demographics, our distribution skew is similar to what is found in existing literature: young, female, and educated [18, 22, 31, 67]. Without controlling for demographics, the proportion of females in the dataset was 79%, those under 30 made up 85% of the overall dataset and consisted of approximately 73% people who claimed white ethnicity. After attempting to boost the under-represented populations, we were able to obtain considerable reductions in proportion of females (73%, proportion of white ethnic background was reduced to 67%, and proportion of those under 30 was reduced to 67%. This increases the average dataset age to 29.5 years old from 25.5 years old. For comparison, the US population is approximately 62% white, 50% female, and 40% under 30. As a large portion of analysis is centered around over-represented versus under-represented groups, we will will frequently refer to under-represented groups as *other* or *non-majority* when presenting results. For example a comparison of female versus non-female or other. We do this to avoid having very small groups due to finer categorization.

Other demographic aspects of the dataset for people who gave answers were: 46% of people identify as monocultural and 41% of people identify as multi-culture for culture identity; 95% were born in the U.S., however only 89% learned English as their first language. Approximately 41% of participants were currently students where 2% of annotators had a doctorate degree, 12% had a graduate degree, 40% had an undergraduate degree, 30% had a high school degree, 14% have a technical degree, and the remainder have other or no formal qualifications. All demographic information and annotations are publicly accessible via Github for further analysis.

### 5.2 Significant effects in sentiment rating due to demographics

We wish to understand whether demographic background can cause significant differences in ratings. For example, differences between older versus younger annotators. To understand the demographic differences while accounting for the various grouping effects from samples, subjects and modalities, we conduct a linear mixed effects analysis following [74]. We construct a linear model of sentiment as a function of all gathered demographics. We modeled age, gender, ethnicity, cultural background, and education. This model was significant ($p < 0.001$). Figure 2 presents




(a) Age Category (b) Simplified Ethnicity (c) Gender

Fig. 2. Mean sentiment ratings broken down by demographic categories. We provide the age, ethnicity and gender charts. Component modalities (text, video, audio, and combined) are shown as different colors. Grey line illustrates 95% confidence intervals. Note that there are obvious significant differences in ratings for modality. Trends can also be observed visually.

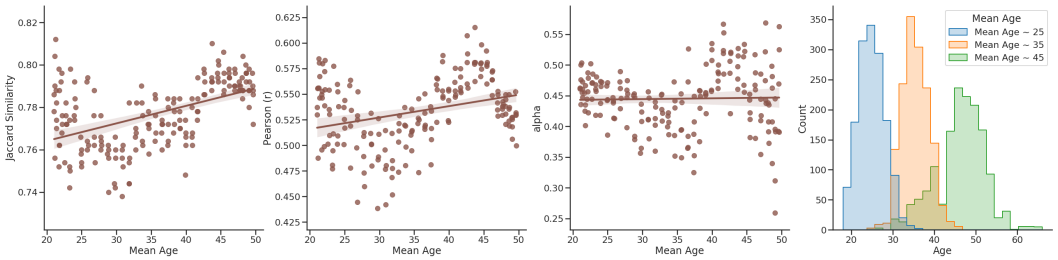
example annotation distributions for age, ethnicity, and gender. We find via visual inspection that despite age having a non-linear effect, the slope is still significant ($p < 0.01$).

We report mixed effects analysis results as an effect size with a standard error bound. We found that age affects sentiment (($\chi^2(1)$=12.17, $p < 0.001$)), and increasing sentiment by approximately $0.0045 \pm 0.0013$ (standard errors) per year. Gender was also found to have a significant effect (($\chi^2(1)$=4.33, $p$=0.037)), increasing rating by $0.066 \pm 0.032$ standard errors. Significant interaction effects were found between gender and ethnicity (($\chi^2(4)$= 11.10, $p$=.026)). Borderline significant interaction effects was found for age and gender (($\chi^2(1)$=3.73, $p$=.053)). We further test for any interaction effects between demographics and modality. As expected, testing modality is a highly significant effect (($\chi^2(1)$=11.10, $p < .001$)). Interaction tests between modality with age, ethnicity, culture and education showed no major effects, the greatest significance was between gender and modality at (($\chi^2(2)$=2.95, $p$=.086)).

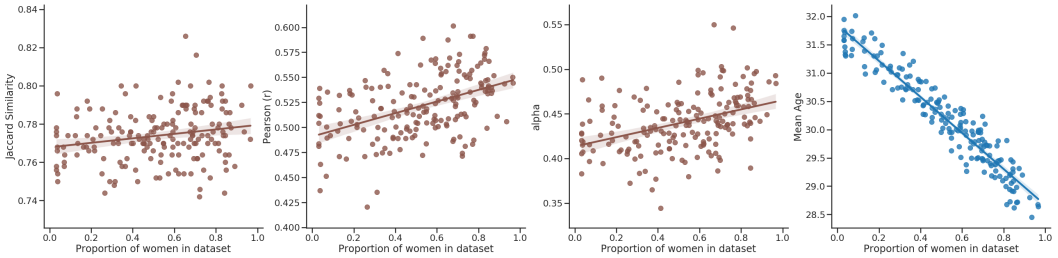
These results suggest that ratings for sentiment when annotating in text, visual, audio and all combined modalities will produce different ratings depending on which demographic is annotating. They also suggest that the demographic effect is consistent across modalities. Furthermore, the significant differences in annotations within each modality suggest that ground truth annotations for each modality is subtly different. That is, human perceptions of text, video, audio, or combined modalities are subtly different. Therefore, when building datasets, we need to be careful whether participants are accounting for the information in the modalities we are interested in holistically. Furthermore, when developing a model for prediction, we should be wary that the predicted label matches the expected label of user interaction. That is, we do not want a model to infer the multimodal sentiment label when a user is only communicating via text, as this might result in lower perceived model effectiveness.

### 5.3 Monte Carlo simulation of demographic effects

We perform a Monte Carlo experiment to visualize how shifting demographics can alter the truth of labels for a dataset. The diverse demographic information in the dataset enables us to perform a Monte Carlo sampling of sub-populations to understand the effect of varying the population




(a) Monte Carlo performance metrics by age. Sampling is performed to obtain a subset of our data to obain differing mean ages. Similarity with original annotations from MOSEI [83] is reported in the left two figures using Jaccard similarity and Pearson's $r$. Agreement within the Monte Carlo sampling is reported as Krippendorf alpha. We report the distribution of demographics in the right-most figure. Observe that when compared to the original dataset (right two figures) that there is a high degree of relative change. In particular, annotators aged approximately 30 had the least amount of similarity with original annotations. Stronger agreement within the dataset (third from right) appears to correspond with more similarity to original annotations.




(b) Monte Carlo performance by gender distribution. From the right two figures, we see that as the proportion of women increases, there is more similarity with the original labels. Since a majority of annotators are women, this shows that the labels are more biased towards the opinions of women. Women also tended to agree more with the annotations of other women. The right most figure demonstrates that there are more younger women than older women and helps to illustrate the co-variance between age and gender.

Fig. 3. Visualization of Monte Carlo experiments on age and gender. Left two figures in each category show similarity with original ratings from [83]. Alpha is the agreement score of the dataset sampled via Monte Carlo. The trends observed above support the significance of effects found in Section 5.2

.

with differences in the dataset. We seek to empirically demonstrate how sampling various sub-populations based on demographic ratios impacts the dataset metrics. We also seek to show how using our labels can be used as an evaluation to test for the spread of performance in current algorithms due to demographic differences. We conduct the results using the combined modality for bench-marking and analysis purposes.

Following the mixed effects analysis, we saw a significance in the annotator gender and age categories for all the modalities. We randomly sample 3 raters per video segment with replacement from the dataset for different mean ages and for varying proportions of women. We examine metrics with respect to the original dataset [83]. We report Jaccard similarity and Pearson correlation against this dataset, similar to how previous models used these [66]. We additionally report the Krippendorf Agreement score for each monte carlo sample.

As can be observed in Figure 3, we observe large variations in age metrics when adjusting for the mean age of the dataset. There appears to be a correlation of within dataset agreement for age

| Age | Text | Video | Audio | All | $r$ |
|---|---|---|---|---|---|
| **$\leq$ 20** | 0.45 | 0.31 | 0.43 | 0.47 | 0.67 |
| **21-30** | 0.47 | 0.33 | 0.43 | 0.48 | 0.71 |
| **31-40** | 0.44 | 0.30 | 0.41 | 0.41 | 0.68 |
| **$>$ 40** | 0.39 | 0.30 | 0.43 | 0.43 | 0.59 |
| **Overall** | 0.45 | 0.32 | 0.42 | 0.46 | 0.75 |

(a) Agreement scores broken by age and modality. Annotators over 40 had slightly lower agreement within themselves. Their predictions also correlated less with previous annotations.

| Ethnicity | Text | Video | Audio | All | $r$ |
|---|---|---|---|---|---|
| **White** | 0.46 | 0.33 | 0.45 | 0.48 | 0.74 |
| **Other** | 0.44 | 0.30 | 0.36 | 0.42 | 0.67 |
| **Overall** | 0.45 | 0.32 | 0.42 | 0.46 | 0.75 |

(b) Agreement scores by ethnicity and modality. Non-white annotators had lower within-group agreement and lower correlation with previous labels. Lower agreement is observed in all modalities.

| Gender | Text | Video | Audio | All | $r$ |
|---|---|---|---|---|---|
| **Female** | 0.48 | 0.32 | 0.44 | 0.47 | 0.76 |
| **Other** | 0.38 | 0.27 | 0.35 | 0.42 | 0.60 |
| **Overall** | 0.45 | 0.32 | 0.42 | 0.46 | 0.75 |

(c) Agreement scores by gender and modality. Non-female annotators had lower within-group agreement and lower correlation with previous labels. Reduction in agreement is observed in all modalities.

Table 3. Agreement scores by modality for age, ethnicity, and gender. For text, video, audio, and all modalities, we report Krippendorff's alpha computed using a variable number of annotators and accounts for missing data. We also report the correlation of the labels (all modalities only) provided by each demographic with the original MOSEI labels ($r$). Older ($>$ 40), non-white, and non-female populations all demonstrated lower agreement with original labels. This further showcases the bias when not controlling for annotator demographics during annotation. The original annotations were obtained via Mechanical Turk which had higher proportions of younger, white and female annotators.

with regards to agreement with previous labels. When controlling for proportion of females, we see a small improvement in metrics as the proportion of female annotators increase. This supports works in literature on the influences of gender in emotion interpretation [3, 12, 50, 51, 57, 73]. Additionally, all these results agree with our observations for significance previously. Furthermore, they support the finding that testing for demographics can be beneficial for measuring ground truth quality in subjectively annotated datasets. In summary, these results suggests that the original ground truth labeled via Mechanical Turk likely follow the overall crowd-working demographic biases, and these effects showcase a demonstrable effect.

### 5.4 Differences in inter-annotator reliability due to demographics

In this section we provide experiments for annotation quality and reliability. We explore two questions: 1) Are annotators in certain demographics more in agreement than others? and 2) Are annotators in certain demographics more in agreement with the original dataset? To evaluate agreement within a demographic group, we use Krippendorf's alpha. Krippendorf's alpha is a metric used to measure annotation consistency among annotators and to give an indication to the quality and amount of variability present in a dataset. It can also normalize for missing data and is applicable on a variable number of coders. To measure agreement with the original dataset, we compute the Pearson correlation ($r$) of each demographic with the original labels from MOSEI.

Correlation is reported for the combined modality only as the original dataset does not provide per-modality annotations. Results are reported in Table 3

The overall krippendorf agreement of our data is .48 which is good for a publicly annotated dataset, especially given the diverse population which provided annotation. Additionally, this is comparable to the .51 reported in [41]. While works such as [59] report higher agreement scores, these annotation efforts typically require a post-annotation discussion phase to find score consensus. It is challenging for crowd-sourced annotated data to do this and thus explains much of this difference. Some works such as [61] have studied how to effectively incorporate a deliberation process into the crowd-sourced annotation process. However, when creating larger datasets, incorporating deliberations have thus far not been used extensively. Potential for future work such as using advanced semi-supervised models on small strongly annotated datasets that incorporate deliberations exist.

We measure the agreement score among sub populations to look for any large demographic effects. No large differences in agreement were noticed in age, with the exception of the text modality and all modalities being slightly lower for annotators over 30. This demonstrates that within each age group, participants had similar opinions regarding the sentiment of a sample. However when examining the correlation, participants over 40 provided annotations that were far less correlated with the original labels. This trend is observed for non-white annotators, as well as non-female annotators. In addition to using the seven-class annotations to compute agreement, we also simplified the labels to be binary and computed the agreement. That is, all labels less than zero are considered to be negative, and all labels greater than zero are considered to be positive. We found that the agreement scores follows a similar trend of higher agreement in over-represented groups and lower agreement in under-represented groups.

These results demonstrate that certain demographic groups might agree on labels more than others. To improve this, some demographic groups may benefit from additional training due to task familiarity. Additional demographic factors such as differences in emotion interpretation due to age, gender, and culture might also be influencing the results. Furthermore, we see that the same groups that have lower agreement (older than 40, non-white, and non-female) also had lower correlation scores with the original annotations. These demographic groups are also less represented among crowdworkers. This is further evidence that the demographics has a strong influence on ground truth labels.

### 5.5 Best/Worst case analysis by demographic

We further quantify the effect of shifting demographics on trained model performance. We experimented with the best possible (and worst possible) demographic distributions with respect to the MOSEI dataset. In other words, what is the population distribution that gives us video labels as close (or as far) as possible to MOSEI? This is relevant because MOSEI and many other sentiment prediction datasets [81] are often taken as ground truth in various works, even though annotator demographics are typically unaccounted for. We wanted to understand the possible swing in scores that could occur with an arbitrarily good (or bad) population distribution.

*5.5.1 Sampling procedure.* To perform this experiment, we first divide the population of annotators into a series of age bins from 18-20, 20-25, ..., 45-50, > 50. We further breakdown the annotators into female or non-female bins. This gives us 16 bins to optimize. Given each bin, we then adjust the weights of the female and non-female annotators to match the desired target gender distribution within that age demographic. The "ground truth" labels for each video sample can then be computed using a weighted averaging of the demographic category based on the mean ratings within each

| Model | $Acc_2$ | F1 | MAE | $r$ |
|---|---|---|---|---|
| **MulT** | **0.778** | 0.779 | 0.724 | 0.617 |
| **MAG-Bert** | **0.756** | 0.752 | 0.698 | 0.679 |
| **MAG-XLNet** | **0.766** | 0.760 | 0.733 | 0.683 |
| **Human** | **0.818** | 0.711 | 0.661 | 0.748 |

(a) Most similar sampling performance metrics.

| Model | $Acc_2$ | F1 | MAE | $r$ |
|---|---|---|---|---|
| **MulT** | **0.729** | 0.734 | 0.905 | 0.506 |
| **MAG-Bert** | **0.715** | 0.714 | 0.879 | 0.551 |
| **MAG-XLNet** | **0.729** | 0.726 | 0.909 | 0.556 |
| **Human** | **0.785** | 0.685 | 0.845 | 0.618 |

(b) Least similar sampling performance metrics.

| Model | $Acc_2$ | F1 | MAE | $r$ |
|---|---|---|---|---|
| **MulT** | 0.770 | 0.771 | 0.708 | 0.623 |
| **MAG-Bert** | 0.756 | 0.752 | 0.678 | 0.690 |
| **MAG-XLNet** | 0.754 | 0.748 | 0.727 | 0.689 |
| **Human** | 0.806 | 0.703 | 0.644 | 0.752 |

(c) US population distribution sampling.

| Model | $Acc_2$ | F1 | MAE | $r$ |
|---|---|---|---|---|
| **MulT** | 0.791 | 0.795 | 0.599 | 0.625 |
| **MAG-Bert** | 0.811 | 0.811 | 0.584 | 0.695 |
| **MAG-XLNet** | 0.861 | 0.860 | 0.551 | 0.746 |
| **Human** | 1.000 | 1.000 | 0.000 | 1.000 |

(d) Comparison with original dataset

Table 4. Performance metrics when measured against different demographic distributions of our annotations. We see some recent models have reduced performance metrics when measured against different sampling techniques. Binary accuracy measured according to [66]. The human model is the original human (MOSEI) annotations compared against our new annotations. The difference w.r.t. original dataset is minimal due to the difference in population and training effect differences. Demographics has a much larger effect on learned models. The difference in performance provides demographic bounds on algorithmic performance. Notice the large difference in accuracy for the same model for most similar and least similar sampling (**bold**). This indicates that demographic differences of annotators can account for more than 4.5 percent difference in performance.

demographic category:

$$v_i = \frac{\Sigma_j w_j l_i}{\Sigma_j w_j}, \tag{1}$$

where $v$ is the determined ground truth label using the simulated distribution, $w_j$ is the weight for the $j$-th group of annotators for the video, $l_i j$ is their label. Weights are optimized via gradient descent using the MAE of our predictions $v_i$ against the original MOSEI labels as loss. By maximizing the MAE with respect to MOSEI, we can obtain the worst case demographic population. We found that the best or worst case population demographics did not change between multiple optimization runs. We restrict the minimum demographic to being 1% of the overall population. For the US population data, we base the demographic weights on census data. We then compare the ratings of this hypothetical population of annotators.

*5.5.2 Models.* Three recent state of the art techniques for multimodal sentiment classification are used for evaluation:

**MulT** [66] is an extension of the transformer architecture to enable multimodal inputs. It incorporates elements of early feature fusion by mixed-attention of modalities and then using late fusion to combine predictions across modalities. We used the unaligned model for evaluation.

**MAG-Bert** [55] Enables the fusion of modalities and the use of pretrained embeddings by exploiting modality gating mechanism inspired by [70] and incorporating into a transformer architecture. State-of-the-art benchmarks were reported on multiple datasets using BERT embeddings. [16]

**MAG-XLNet** Utilizes the same fusion technique however uses an XLNet [80] backbone which is improvement on Bert that exploits autoregressive training, relative positioning, and segment recurrence from Transfomer-XL[13] for improved modeling.

Each algorithm is trained on the original dataset using publicly available code. The algorithm results are then measured against the ratings scores for a particular population distribution derived from the optimization procedure. We report the binary accuracy, F1 score, mean average error and pearson correlation. For binary accuracy, this value is reported as the accuracy of positive or negative sentiment only. F1 is a harmonic mean of precision recall for positive and negative sentiment. MAE and correlation is a measure obtained from the means for a specific sample and the predicted mean rating. Note also that since our annotations are for the test set, we run for only a single trial as opposed to cross-validation. This methodology is the same as existing standards.

*5.5.3 Classification results.* We present the results in Table 4. As can be seen, there is a large spread in performance among different population distributions. A drop in performance is expected as the existing works do not optimize for our test condition. The most similar sampling to original labels did not improve results significantly, as expected, potentially due to crowd-working demographics being similar. Examining the US population shows that there is a drop, indicating that the demographic differences are having an effect of about 1%, when compared against a "crowdworker" demographic.

However, what is surprising is the potential effect from a demographic that is the least similar to the original annotations. While we see an approximately 3.3% drop in binary accuracy between least and most similar for when compared to the original dataset, algorithmic performance decreased much more. Algorithm performance decreased up to 4.9%. This difference suggests that the algorithm is over-fitting to properties in the original dataset, and that these properties can be observed when adjusting for demographics. From an HCI perspective, the effect of this would suggest that AI systems for sentiment prediction works well for some people (e.g., younger, white, and female) and not others (e.g., older, non-white, and non-female).

Furthermore, when measured against the original annotations (Table 4(d)), the new models MAG-Bert and MAG-XLNet outperform the older technique (MulT). However, when the population is changed, we see that newer techniques perform much worse than older techniques. This suggests that the newer models are matching patterns and properties in the original dataset that can be quantified via annotator demographics.

In summary, we see that by sampling for different demographics we can place a bound on the expected behavior due to variations in annotator demographics. From our experiments, this effect is almost 5% for *binary* accuracy and affects models differently. This is quite significant as we are simply evaluating based on positive and negative expressed sentiment and not at a fine grained level. For more recent models (MAG-XLNET and MAG-Bert) the performance drops *more* than the older technique (MulT). As models have become more capable of capturing dataset nuances, these effects appear to become more amplified based on this experiment. This points to the importance for more rigorous evaluation measures that include annotator demographic information. This is particularly important for annotations that have high degrees of subjectivity.

## 6 DISCUSSION

We discuss the impact of annotator demographics on dataset biases and model efficacy, current limitations of our work, and recommendations for the research field.

| Model | $Acc_2$ | F1 | MAE | $r$ |
|---|---|---|---|---|
| **MulT** | 0.756 | 0.759 | 0.746 | 0.604 |
| **MAG-Bert** | 0.746 | 0.742 | 0.714 | 0.666 |
| **MAG-XLNet** | 0.754 | 0.752 | 0.755 | 0.664 |
| **Human** | 0.808 | 0.712 | 0.673 | 0.733 |

(a) Comparison to younger annotators (<30).

| Model | $Acc_2$ | F1 | MAE | $r$ |
|---|---|---|---|---|
| **MulT** | 0.772 | 0.771 | 0.785 | 0.591 |
| **MAG-Bert** | 0.748 | 0.741 | 0.810 | 0.643 |
| **MAG-XLNet** | 0.756 | 0.747 | 0.763 | 0.651 |
| **Human** | 0.792 | 0.707 | 0.749 | 0.691 |

(b) Comparison to older annotators ($\geq$ 30)

| Model | $Acc_2$ | F1 | MAE | $r$ |
|---|---|---|---|---|
| **MulT** | 0.768 | 0.771 | 0.737 | 0.616 |
| **MAG-Bert** | 0.746 | 0.742 | 0.705 | 0.681 |
| **MAG-XLNet** | 0.754 | 0.751 | 0.753 | 0.674 |
| **Human** | **0.816** | 0.711 | 0.652 | 0.755 |

(c) Comparison to female annotators.

| Model | $Acc_2$ | F1 | MAE | $r$ |
|---|---|---|---|---|
| **MulT** | 0.732 | 0.736 | 0.882 | 0.533 |
| **MAG-Bert** | 0.742 | 0.740 | 0.833 | 0.605 |
| **MAG-XLNet** | 0.726 | 0.721 | 0.868 | 0.609 |
| **Human** | **0.760** | 0.693 | 0.833 | 0.621 |

(d) Comparison to non-female annotators.

| Model | $Acc_2$ | F1 | MAE | $r$ |
|---|---|---|---|---|
| **MulT** | 0.766 | 0.766 | 0.750 | 0.606 |
| **MAG-Bert** | 0.752 | 0.747 | 0.704 | 0.678 |
| **MAG-XLNet** | 0.752 | 0.749 | 0.759 | 0.670 |
| **Human** | 0.798 | 0.704 | 0.678 | 0.738 |

(e) Comparison to white annotators.

| Model | $Acc_2$ | F1 | MAE | $r$ |
|---|---|---|---|---|
| **MulT** | 0.762 | 0.764 | 0.809 | 0.567 |
| **MAG-Bert** | 0.756 | 0.753 | 0.789 | 0.616 |
| **MAG-XLNet** | 0.748 | 0.746 | 0.821 | 0.628 |
| **Human** | 0.774 | 0.698 | 0.761 | 0.680 |

(f) Comparison to non-white annotators.

Table 5. Performance metrics when measured against different demographic distributions of our annotations. We compare against specific demographics to observe differences for certain user groups. Assuming that users of a certain demographic prefer annotations from the same demographic, we can see a difference in performance for different user groups. Tables on left (a, c, e) present results that represents the majority of crowdworker demographics. We can see our scores from demographics which match the majority demographic population far better than those that match the minority. Reduction in model performance is also pronounced for female vs non-female annotators (c vs d). The trend of decrease between majority vs minority demographics is obvious via MAE and correlation.

### 6.1 Impact

In this work we produced a set of annotations large enough for machine learning evaluations that contains detailed demographic information. We find that there can be a nearly 5% difference (77.8%-72.9% in the case of MulT) in binary classification accuracy alone when adjusting demographics for evaluating model behavior. This difference is likely exacerbated when when examining fine-grained sentiment classification or for more controversial annotation tasks. For example, in applications such as language toxicity classification, it has been observed that real world user experience and reported algorithmic performance is vastly different [29]. Furthermore, as models are becoming increasingly expressive and optimized with respect to original datasets (which clearly have their own demographic biases), demographic differences may make the seeming improvements much less pronounced and impressive. This can be observed in Table 4 where MulT, an older algorithm, outperforms the newer algorithms MAG-Bert and MAG-XLNet when labels are given by annotators which follow a US population distribution. We further contextualize this expectation via hypothetical user groups who use sentiment prediction algorithms and frameworks in Table 5. With no exception, we see that demographics that correspond to the majority class correlate better with previous

annotations. These results mean that any user who belongs to a minority demographic group (with respect to the overall annotator demographics distribution) will perceive the sentiment rating system to perform worse than those in the majority group. Given that there are quite different biases in common machine learning corpora, namely biases towards white male populations, and common crowdsourcing populations, namely biases towards young white female populations – both uniquely problematic —, this situation is bound to occur fairly often.

Many techniques have tackled the issue of mismatch in real world versus experimental metrics from different angles. We find that annotator biases quantified by demographics might be one important source of the issue. This would suggest that existing datasets, while valuable and necessary for the development of learning models, do not work well for a large portion of the population in practice. However, as previous works have pointed out, biases can be removed by increasing the number of annotators in diverse groups [32]. This suggest that one solution might be to extend current datasets with additional annotations from less represented demographics.

### 6.2 Limitations

While we see the data and analysis as highly beneficial for the domain, there are limitations to the answers that our work can provide. The data gathered was limited to a single country (US), and more work is needed to understand the effect of a wider demographic on machine labeled data. It may be for this reason that we did not see a significant effect for ethnicity, and culture across regions could change (likely amplify) the significance of certain effects. Additional effort will need to be made to examine the differences across cultures and the effects. Another demographic issue is that while we obtained data for additional gender categories, we could not obtain sufficient amounts of data to model gender as a non-binary demographic. For this reason, we resorted to analysis using a proportion of women in the overall dataset. This allowed us to show that differences do exist when accounting for gender and the degree of this effect. In our study, we compromised on these demographic choices as including them would have drastically exacerbated the long-tailed distribution of crowdworker demographics. The availability of annotators from certain demographic groups was frequently very low.

Considerations regarding dataset type should also be made. The annotations are for a specific kind of data – opinions and monologue videos. While it is an important problem, there is a lot more to sentiment recognition research beyond just talking heads and opinions. The MOSEI dataset is mostly comprised of video samples which is less controversial than topics such as hate speech [2, 26, 40, 72]. While previous works have demonstrated that demographic imparts differences in ratings for videos of differing content type, the degree of this effect does not appear to be quantified. Furthermore, labeling content that is less subjective might also demonstrate different effects. For example, in the case of determining dogs versus cats, demographic background likely play a smaller role. We provide our rich annotations for future researchers to understand and correctly these differences.

### 6.3 Recommendations

We echo existing calls for caution when using ML systems. We recommend that ML practitioners should be cautious when implementing technologies that use sentiment prediction models and that users should take great care in interpreting model predictions. This is particularly important in high risk scenarios such as in clinical settings. We hope that current users and practitioners can use our results to interpret ML predictions in a new light. In our experiments, we found that the attributes of annotators imparted a significant difference on both the ground truth as well as the model predictions. One potential approach which follows from this is to match the demographic properties of the annotator with the users of the predictions of the models. Or, more generally, we can

try to match annotator distribution with user distribution to maintain performance of any system. Because naturally, demographics do not explain all of the variability or distribution differences in annotator properties. However, by restricting the user distribution, HCI researchers are greatly restricting themselves in the experiments they can conduct and the designs they can create. As models become more capable, data variability, such as those arising due to demographics, is more readily captured by the model. Yet at the same time, without understanding what this variability is, it can be difficult to both improve the model or interpret the results correctly. For example, suppose we did not know that most dataset annotators are female, then we would be confused as to why a sentiment prediction software works better for female than for non-female users. This highlights an important need for increased collaboration between ML and HCI researchers to develop better models and to build more representative datasets.

We recommend the collection and release of properly anonymized annotator demographic information for subjective tasks such as sentiment or emotion labeling. This recommendation was also voiced by previous work (e.g. [53]), showing significance on similar tasks. Our analysis provides strong evidence that machine learning researchers in particular need to be mindful of the demographic composition of human annotators. As added evidence for the importance of this, we show that the effect on algorithms is larger than the expected effect when comparing different human annotations. Release of our data will facilitate the development of improved algorithms for predicting distributions of multimodal sentiment classification for different demographic groups. This would lead to the improved experience of users which are in different demographic groups than the majority of those who provided annotations.

We recommend the development of a richly annotated subset of data to help quantify variability or annotation noise. In this work, we examined the effect of demographics and we suspect that future work will demonstrate other biases that occur on different dimensions. Annotating a subset of data to quantify the degree of variability due to biases is sufficient for analysis and is also considerably more cost-effective than duplication of entire datasets with additional demographic information. The annotation effort for this work for 500 samples for all modalities cost approximately 3000 USD, or approximately 750 USD per modality. We find this cost to be reasonable for any large scale annotation effort. Understanding dataset biases in this manner can substantially benefit future users from diverse groups, and the insights can likely be transferred to larger datasets.

Lastly, we recommend the balancing of demographic backgrounds of annotators during dataset creation. In our experiments, we found that certain demographic effects were amplified potentially due to the task being more appealing to certain populations. And one has to be mindful of the danger of oversampling individual annotators in highly underrepresented demographic groups. We took a less aggressive approach with regard to enforcing parity of underrepresented demographic groups and saw a benefit to the overall representation. Such a method is easy to implement in practice and can potentially be combined with methods such as [32]. While not perfect in that it would not result in the ultimately desired (e.g. uniform) distribution, the improved representation might increase the benefit to more groups of future users, and statistical methods, paired with the insights from papers such as ours, can be used to approximate the desired distributions.

## 7 CONCLUSION

The goal of this work was to gain an understanding for the variability that subjectively annotated datasets might contain. Towards this goal, we present a large scale dataset that captures annotator demographics variability and contains annotations for multimodal data and its component modalities. We demonstrate the importance for understanding annotator demographics. We show that that demographic differences impute a significant effect on the ratings they provide and that these

effects occur in all modalities. We verify these properties and show large algorithmic performance variability when measured against different demographic groups.

As models become more complex and capable of modeling the details of human expression, more thorough evaluations which can account for the biases in data should be conducted. As is the case here, models and evaluations weigh the opinions of those who perform crowdwork versus those who do not differently. This leaves the potential to bias evaluations and model selection to people who are not part of the annotation process. We hope our data and results can be beneficial not only for future researchers who wish to build more representative datasets, but for evaluation of algorithms and understanding annotator behavior.

## 8 ACKNOWLEDGEMENTS

We would like to thank Lina Kim for her support via the Research Mentorship Program at UC Santa Barbara. This work was funded in part by ONR awards N00014-19-1-2553 and N00174-19-1-0024, as well as NSF award IIS-1911230.

## REFERENCES

[1] Mayumi Adachi, Sandra E Trehub, and Jun-Ichi Abe. 2004. Perceiving emotion in children's songs across age and culture 1. *Japanese Psychological Research* 46, 4 (2004), 322–336.

[2] Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 184–190.

[3] Sara B Algoe, Brenda N Buswell, and John D DeLamater. 2000. Gender and job status as contextual cues for the interpretation of facial expression of emotion. *Sex roles* 42, 3 (2000), 183–208.

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[5] Ruha Benjamin. 2019. Race after technology: Abolitionist tools for the new jim code. *Social forces* (2019).

[6] Md Momen Bhuiyan, Amy X Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.

[7] Leslie R Brody. 1997. Gender and emotion: Beyond stereotypes. *Journal of Social issues* 53, 2 (1997), 369–393.

[8] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335–359.

[9] Quan Ze Chen, Daniel S Weld, and Amy X Zhang. 2021. Goldilocks: Consistent Crowdsourced Scalar Annotations with Relative Uncertainty. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.

[10] John Joon Young Chung, Jean Y Song, Sindhu Kutty, Sungsoo Hong, Juho Kim, and Walter S Lasecki. 2019. Efficient elicitation approaches to estimate collective crowd answers. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[11] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 4067–4080.

[12] John Condry and Sandra Condry. 1976. Sex differences: A study of the eye of the beholder. *Child development* (1976), 812–819.

[13] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860* (2019).

[14] Elizabeth Davis, Ellen Greenberger, Susan Charles, Chuansheng Chen, Libo Zhao, and Qi Dong. 2012. Emotion experience and regulation in China and the United States: how do culture and gender shape emotion responding? *International Journal of Psychology* 47, 3 (2012), 230–239.

[15] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. (July 2020), 4040–4054.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[17] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[18] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 135–143.

[19] Yi Ding, Brandon Huynh, Aiwen Xu, Tom Bullock, Hubert Cecotti, Matthew Turk, Barry Giesbrecht, and Tobias Höllerer. 2019. Multimodal Classification of EEG During Physical Activity. In *2019 International Conference on Multimodal Interaction*. 185–194.

[20] Yi Ding, Alex Rich, Mason Wang, Noah Stier, Pradeep Sen, Matthew Turk, and Tobias Höllerer. 2021. Sparse Fusion for Multimodal Transformers. *arXiv preprint arXiv:2111.11992* (2021).

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR* abs/2010.11929 (2020).

[22] James Feyrer. 2007. Demographics and productivity. *The Review of Economics and Statistics* 89, 1 (2007), 100–109.

[23] Agneta H Fischer, Patricia M Rodriguez Mosquera, Annelies EM Van Vianen, and Antony SR Manstead. 2004. Gender and culture differences in emotion. *Emotion* 4, 1 (2004), 87.

[24] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. 2017. Modus operandi of crowd workers: The invisible role of microtask work environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–29.

[25] Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. 2017. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 4 (2017), 1–26.

[26] Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898* (2019).

[27] Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *proceedings of the 2018 conference on empirical methods in natural language processing*. 3454–3466.

[28] Chris Giordano, Meghan Brennan, Basma Mohamed, Parisa Rashidi, François Modave, and Patrick Tighe. 2021. Accessing Artificial Intelligence for Clinical Decision-Making. *Frontiers in Digital Health* 3 (2021), 65.

[29] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 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*. 1–14.

[30] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[31] Jeff Howe et al. 2006. The rise of crowdsourcing. *Wired magazine* 14, 6 (2006), 1–4.

[32] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[33] David R Karger, Sewoong Oh, and Devavrat Shah. 2011. Iterative learning for reliable crowdsourcing systems. Neural Information Processing Systems.

[34] Dacher Keltner, Deborah H Gruenfeld, and Cameron Anderson. 2003. Power, approach, and inhibition. *Psychological review* 110, 2 (2003), 265.

[35] Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2018. The Gab Hate Corpus: A collection of 27k posts annotated for hate speech. (2018).

[36] Eugenia Kim, De'Aira Bryant, Deepak Srikanth, and Ayanna Howard. 2021. Age bias in emotion detection: an analysis of facial emotion recognition performance on young, middle-aged, and older adults. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 638–644.

[37] Chu Kim-Prieto and Ed Diener. 2009. Religion as a source of variation in the experience of positive and negative emotions. *The Journal of Positive Psychology* 4, 6 (2009), 447–460.

[38] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2011), 18–31.

[39] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Bjoern W Schuller, et al. 2019. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE transactions on pattern analysis and machine intelligence* (2019).

[40] Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering Annotator Disagreement about Racist Language: Noise or Signal?. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*. 81–90.

[41] Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2018. Computational modeling of human multimodal language: The mosei dataset and interpretable dynamic fusion. In *First Workshop and Grand Challenge on Computational Modeling of Human Multimodal Language*.

[42] Paweł Łowicki, Marcin Zajenkowski, and Patty Van Cappellen. 2020. It's the heart that matters: The relationships among cognitive mentalizing ability, emotional empathy, and religiosity. *Personality and Individual Differences* 161 (2020), 109976.

[43] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* 44, 1 (2012), 1–23.

[44] Winter Mason and Duncan J Watts. 2009. Financial incentives and the" performance of crowds". In *Proceedings of the ACM SIGKDD workshop on human computation*. 77–85.

[45] Rachel LC Mitchell, Rachel A Kingston, and Sofia L Barbosa Bouças. 2011. The specificity of age-related decline in interpretation of emotion cues from prosody. *Psychology and aging* 26, 2 (2011), 406.

[46] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards Multimodal Sentiment Analysis: Harvesting Opinions from The Web. Alicante, Spain.

[47] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention Bottlenecks for Multimodal Fusion. *arXiv preprint arXiv:2107.00135* (2021).

[48] Kento Nishi, Yi Ding, Alex Rich, and Tobias Hollerer. 2021. Augmentation strategies for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8022–8031.

[49] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 973–982.

[50] E Ashby Plant, Janet Shibley Hyde, Dacher Keltner, and Patricia G Devine. 2000. The gender stereotyping of emotions. *Psychology of Women Quarterly* 24, 1 (2000), 81–92.

[51] E Ashby Plant, Kristen C Kling, and Ginny L Smith. 2004. The influence of gender and social role on the interpretation of facial expressions. *Sex roles* 51, 3 (2004), 187–196.

[52] Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology* 17, 3 (2005), 715–734.

[53] Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Díaz. 2021. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699* (2021).

[54] Jordi Quoidbach, Elizabeth W Dunn, Konstantin V Petrides, and Moïra Mikolajczak. 2010. Money giveth, money taketh away: The dual effect of wealth on happiness. *Psychological science* 21, 6 (2010), 759–763.

[55] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2020. NIH Public Access, 2359.

[56] Michaela Riediger, Manuel C Voelkle, Natalie C Ebner, and Ulman Lindenberger. 2011. Beyond "happy, angry, or sad?": Age-of-poser and age-of-rater effects on multi-dimensional emotion perception. *Cognition & emotion* 25, 6 (2011), 968–982.

[57] Michael D Robinson, Joel T Johnson, and Stephanie A Shields. 1998. The gender heuristic and the database: Factors affecting the perception of gender-related differences in the experience and display of emotions. *Basic and Applied Social Psychology* 20, 3 (1998), 206–219.

[58] Ted Ruffman, Julie D Henry, Vicki Livingstone, and Louise H Phillips. 2008. A meta-analytic review of emotion recognition and aging: Implications for neuropsychological models of aging. *Neuroscience & Biobehavioral Reviews* 32, 4 (2008), 863–881.

[59] Koustuv Saha, Asra Yousuf, Louis Hickman, Pranshu Gupta, Louis Tay, and Munmun De Choudhury. 2021. A social media study on demographic differences in perceived job satisfaction. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–29.

[60] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. *arXiv preprint arXiv:2111.07997* (2021).

[61] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–19.

[62] Morgan Klaus Scheuerman, Aaron Jiang, Katta Spiel, and Jed R Brubaker. 2021. Revisiting Gendered Web Forms: An Evaluation of Gender Inputs with (Non-) Binary People. In *Proceedings of the 2021 CHI Conference on Human Factors in*

*Computing Systems*. 1–18.
[63] Aaron D Shaw, John J Horton, and Daniel L Chen. 2011. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. 275–284.
[64] Sara E Snodgrass. 1985. Women's intuition: The effect of subordinate role on interpersonal sensitivity. *Journal of Personality and Social Psychology* 49, 1 (1985), 146.
[65] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*. 254–263.
[66] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558.
[67] Stephen Uzor, Jason T Jacques, John J Dudley, and Per Ola Kristensson. 2021. Investigating the Accessibility of Crowdwork Tasks on Mechanical Turk. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
[69] Liuping Wang, Dakuo Wang, Feng Tian, Zhenhui Peng, Xiangmin Fan, Zhan Zhang, Mo Yu, Xiaojuan Ma, and Hongan Wang. 2021. Cass: Towards building a social-support chatbot for online health community. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–31.
[70] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7216–7223.
[71] Fabian L Wauthier and Michael Jordan. 2011. Bayesian bias mitigation for crowdsourcing. *Advances in neural information processing systems* 24 (2011), 1800–1808.
[72] Maximilian Wich, Hala Al Kuwatly, and Georg Groh. 2020. Investigating annotator bias with a graph-based approach. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 191–199.
[73] Sherri C Widen and James A Russell. 2002. Gender and preschoolers' perception of emotion. *Merrill-Palmer Quarterly (1982-)* (2002), 248–262.
[74] Bodo Winter. 2013. Linear models and linear mixed effects models in R with linguistic applications. *arXiv preprint arXiv:1308.5499* (2013).
[75] Christine Wolf and Jeanette Blomberg. 2019. Evaluating the promise of human-algorithm collaborations in everyday work practices. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.
[76] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. 2020. Investigating bias and fairness in facial expression recognition. In *European Conference on Computer Vision*. Springer, 506–523.
[77] Jing Nathan Yan, Ziwei Gu, Hubert Lin, and Jeffrey M Rzeszotarski. 2020. Silva: Interactively Assessing Machine Learning Fairness Using Causality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[78] Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 2019. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the third workshop on abusive language online*. 11–18.
[79] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 547–558.
[80] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).
[81] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3718–3727.
[82] Amir Zadeh, Yan Sheng Cao, Simon Hessner, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. 2020. CMU-MOSEAS: A multimodal language dataset for Spanish, Portuguese, German and French. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, Vol. 2020. NIH Public Access, 1801.
[83] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 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.

Received January 2022; revised April 2022; accepted August 2022