Unintentional affective priming during labeling may bias labels

Judy Hanwen Shen, Agata Lapedriza, and Rosalind W. Picard
Massachusetts Institute of Technology
Cambridge, USA
{judyshen, agata, picard}@mit.edu

***Abstract*—Online platforms displaying long streams of examples are often employed to gather labels from both experts and crowd workers. While previous work in crowdsourcing focused on objective tasks and estimating error parameters of annotators, collecting labels in a subjective setting (e.g. emotion recognition) is more complicated due to different interpretations of examples. These interpretations could be influenced by many factors such as annotator mood and previously seen examples. In this work, we examine two hypotheses of order-dependent biases in sequential labeling tasks: negatively auto-correlated sequential decision making and positively auto-correlated affective priming. Using controlled generation of facial expressions, we find that i) annotators achieve higher agreement when presented examples in the same sequential order, ii) the valence label of the current image positively correlates with the previous labels given. While we also observe a positive correlation between labels and the number of preceding positive and negative images seen, this correlation is highly dependent on example ordering. Our findings demonstrate that randomized examples given to annotators may produce systematic bias in labels. Future data collection should present examples in orderings which mitigate such bias.**

***Index Terms*—Affective computing, emotion recognition, computer vision, crowdsourcing**

I. Introduction

Building intelligent models for subjective tasks such as sentiment analysis and emotion recognition is of interest for downstream tasks such as news recommendations and human-robot interaction. Techniques for aggregating opinions and expertise to produce labels for training have been widely studied [1, 2]. Most crowdsourcing methods assume an objective true label and infer error rates of annotators [3, 4]. However, crowdsourcing for tasks that are subjective in nature can yield noisy labels that arise due to a combination of annotator error and subjectivity. For example, the facial emotion could be perceived differently due to the mood and fatigue of an annotator, and previous examples seen by an annotator. Existing crowdsourcing models do not provide mechanisms to disentangle the effect of multiple opinions from the effect of annotator error in such tasks.

In the popular computer vision task of emotion recognition from facial expressions, large datasets are often labeled by few annotators [5, 6, 7]. This results in an experimental setup where each annotator sees a long sequence of examples (e.g. $36,000$ examples by 2 annotators [5]; $24,000$ examples by 2 annotators [6]). This is a sequential decision-making task that is highly susceptible to systematic label noise. Wisdom from affect priming suggests that annotators may rate the valence of a current example to be similar to that of the preceding example [8]. Conversely, an annotator's belief in randomness (gambler's fallacy or law of small numbers) may lead to negatively auto-correlated labels even among experts [9].

Motivated by the understudied problem of crowdsourcing in affective computing, this work investigates systematic noise in emotion recognition labels which arise due to the subjective nature of the task and the limited quality of crowd workers annotations. We use a state-of-the-art generative model to produce a set of linearly interpolated facial expressions of various emotions. We measure the effect of example ordering on the output label produced by annotators and find that emotion annotation is dominated by affective priming rather than gambler's fallacy. Specifically we investigate the following three questions in the context of annotator behavior in emotion recognition tasks:

- Q1: Is there a difference in inter-annotator agreement between a uniform sequence of images and the randomized sequences of the same set of images?
- Q2: In randomly shuffled sequences, does the valence-based ordering of examples relate to the valence label produced by annotators?
- Q3: What is the effect of the valence of previous images seen by annotators on the label annotators will give to a current neutral image?

II. Related Work

A. Affective Crowdsourcing

The aim of modeling multiple annotators is to collect the most accurate label despite disagreement between annotators. Crowdsourcing techniques have been applied to various affective computing tasks and datasets [10, 11]. Snow et al. [2] asked Amazon's Mechanical Turk workers to label the affect of news headlines. In emotional speech recognition: Parthasarathy and Busso [12] applied preference-learning to generate labels based on changes in emotion that annotators agree on and Lotfian and Busso [13] formulated emotional perception as a multidimensional Gaussian where each dimension corresponds to an emotion.

Sequential ordering in crowd sourcing has rarely been explored although assimilation and contrast effects have been examined in social psychology [14]. Prior works have focused

978-1-7281-3888-6/19/$31.00

Fig. 1. Example of generated expressions at various coefficients for SADNESS to HAPPINESS variation (a) and ANGER to HAPPINESS variation (b). Both ANGER and SADNESS are considered negative valence photos in our generated dataset

on predicting annotator error over time [15, 16]. Atcheson et al. [17] found that perceived emotion in continuous speech can exhibit an inherent degree of ambiguity independent from annotator error. They further find that human ratings for continuous emotion in speech exhibit locally smooth properties. This benefit of prior context in improving labels has also been demonstrated in multi-modal affective tasks [18]. Unlike speech or video windows in continuous emotion recognition, the sequential effects we explore do not have inherent time-dependent properties. For example, a sequence of positive images to be labeled can occur as a result of random shuffling of images while a sequence of positive valence snippets of speech occurs due to speech or video windows being from the same utterance. While the previous information in the same sentence serves as context for the current label in continuous speech emotion annotations, the effect of the previous randomly shuffled images before a current image is usually considered noise to be eliminated.

B. Sequential Decision-Making: "Gambler's Fallacy"

If providing labels for a sequence of examples is modeled as a decision-making problem, previous work in behaviour economics suggests that annotators provide negatively auto-correlated labels due to a belief that sequences of identically labeled examples are improbable. This phenomenon, the gambler's fallacy, occurs in decisions made by asylum judges,loan officers, and baseball empires [9]. Unlike contrast effects, this fallacy arises due to the immediate preceding label given rather than due to stimuli appearing earlier in the sequence [19]. This linear auto-correlation model for observation data can be written as:

$$Y_i = \beta_0 + \beta_1 Y_{i-1} + Controls + \epsilon_i. \quad (1)$$

Chen et al. [9] found that β_1 is negative which allows us to write our hypothesis of this effect as:

$$p(y_i = c) < p(y_i = c | y_{i-1} = c) \quad (2)$$

where y_i is the class label of the i^{th} example and c is a class which examples can belong to. In the setting of labeling affective examples, annotators may assume images appear in random sequences. Extending to k previous images and considering that the current example is similar to previous examples, an annotator may decide that the x_i is less likely to be in class c when x_{i-k} ... x_{i-1} appear to be in class c:

$$p(y_i = c | x_i) < p(y_i = c | x_i, y_{i-1}, ...y_{i-k} = c) \quad (3)$$

C. Affective Priming

A contrasting phenomena of interfacing with sequential examples is affective priming [8]. In affect labeling or perception, a target may be more likely to be of a certain valence category if the previous or primer example is of the same valence category. Experiments confirming affective priming include facial expression perception and image aesthetic perception. In this case, the inequality in equation 4 would be reversed.

$$p(y_i = c | x_i) > p(y_i = c | x_i, y_{i-1}, ...y_{i-k} = c) \quad (4)$$

Leopold et al. [20] showed that there are robust aftereffects of facial perception by testing facial identity recognition. Similar work stretches beyond perception of human faces to the perception of aesthetically pleasing images. Chang et al. [21] performed various experiments which showed that the preference ratings for a neutral image are influenced by whether the previous image is a preferable or less preferable image. Priming can be used to induce better performance among crowd workers. Morris et al. [10] used photos with positive affect to prime crowd workers on Amazon Mechanical Turk to generate more creative responses. With respect to perception of facial expressions in particular, Bouhuys et al. [22] used music to induce depressed and elated moods and tested whether these induced moods affect perception of faces. They found that subjects perceived more sadness in faces showing a preponderance of positive or negative emotions when feeling more depressed. Prior work studying the role of induced emotion on perception in conjunction with work illustrating the role of affective response to facial expressions motivates our work in investigating whether such phenomena also occur in affective labeling tasks.

III. Facial Expression Generation

To test the effect of sequential ordering of affect labels, ground truth images were required. However, collecting "gold-standard" labels for existing emotion recognition datasets would introduce biases this work is trying to address. We thus combine recent work in style transfer using generative adversarial networks (GAN) [23] with posed emotions from the Karolinska Directed Emotional Faces (KDEF) dataset [24] to generate interpolated emotional expressions. The KDEF dataset contains 7 posed emotions from 70 Caucasian actors and actresses. We employ style transfer of high level features using the StyleGAN model [25]. The StyleGAN architecture employs two latent spaces which allow the preservation of features when linearly traversing the second order latent space. We leverage this to generate varying degrees of each emotion for the posed emotions set. For simplicity, we examine the task of collecting 1-dimensional valence labels for emotion recognition. Using HAPPINESS, SADNESS, ANGER, and NO EMOTION poses, we find the latent space representation that minimizes the reconstruction loss using the StyleGAN encoder. With the latent vectors of each emotion, we interpolate to create [1/3, 2/3] interpolations of the emotions displayed in the photos. This generates two variations of 7 degrees of negative to positive valence photos: an ANGER to HAPPINESS variation, and a SADNESS to HAPPINESS variation (Figure 1). The images were then manually filtered for quality, leaving 64 sets of photos (i.e. photos of 64 individuals) for each emotion category. This ensured that images containing undesirable artifacts of image generation such as hair blending into background or blurred spots were removed.

IV. Experiments

Our objective is to closely replicate the environment annotators experience when annotating large emotion recognition datasets. Corresponding to our generated images, we provide 7 valence labels from extremely negative to extremely positive. Figure 2 shows an example question shown to annotators. Annotators were presented sets of 40 (36 for later sequence experiments) images to label. Before each set of images, participants were asked their current mood on the same scale from extremely negative to extremely positive. After each set of images, participants were given a 30-second break to minimize possible affective influence from the previous set. Participants were crowd workers hired through Amazon Mechanical Turk (AMT) and compensated around 10 USD per hour for tasks lasting between 7 and 20 minutes. We also collected the demographic information (e.g. age, race, gender, geographic location) from the crowd workers to assess the composition of annotators since all faces in our dataset are Caucasian. While the crowd workers were a mix of age and genders, most participating workers responded as Caucasian (65%), and from the United States (94%).

Experiment 1: Annotator Agreement

We first examine the role of previous images seen on annotator agreement. We test whether randomizing the order

Fig. 2. Example survey interface shown to annotators

of examples in emotion recognition induces a different level of inter-annotator agreement than when all annotators see the same examples in the same order. This is important to examine in subjective tasks because it is difficult to disambiguate whether annotator disagreement arises from the inherent example ambiguity or from ordering effects.

With a set of sampled generated images, 49 annotators each labeled 80 images (2 sets: 40 images per set) in different orders. Each set of images contained approximately the same ratio of HAPPINESS, SADNESS, ANGER, and NO EMOTION poses. Half of annotators saw the first set of images in the same order and the second set in one of 5 different orderings. The other half of annotators saw the first set of images in one of 5 different orderings and the second set in the same order. Each set of 40 images took around 3-4 minutes to annotate.

To compare the disagreement of annotators between the two orderings of the same set of images, we calculate the entropy of the set of answers for image j as:

$$H_j = -\sum_{i=1}^{n} p_j(a_i) \log p_j(a_i) \tag{5}$$

where $p_j(a_i)$ is the probability of answer a_i for the j^{th} question. Here, $n = 7$ since there are 7 possible answers each annotator could give for each question (Figure 2).

Figure 3 shows the difference in distribution between the entropy of the uniformly-ordered set compared to the shuffled set. The mean of the shuffled set is 1.10 while the mean entropy of the uniformly-ordered set is 1.06. While the difference in means is not statistically significant, figure 3 shows the distributional differences in entropy between the two sets. We also compute the Fleiss' kappa relative to chance agreement for each set. Confirming the entropy distribution

results, the uniformly-ordered set has a higher kappa for annotator agreement ($\kappa = 0.250$) compared to the shuffled set ($\kappa = 0.219$). Here, a higher kappa reflects a higher inter-annotator agreement. As more permutations are labeled by annotators, towards the limit of each annotator seeing a different ordering, we would expect the entropy of answers to increase and the Fleiss' kappa to decrease.

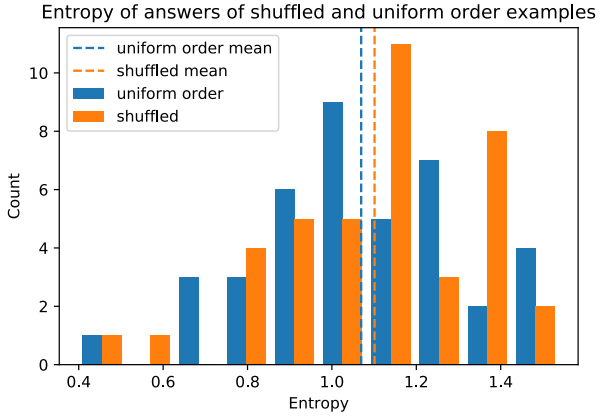

Fig. 3. Entropy of answers in uniformly ordered set (blue) and 5-fold shuffled set (blue). The mean entropy is indicated with dotted lines

Experiment 2: Ordering effects in Random Set

We examine what type of ordering effects exist to cause higher entropy in amassing labels for differently ordered examples than uniformly ordered examples. In the randomly generated blocks, we count the number of previously seen positive and negative images to discern whether there is a correlation between previous images and a current label.

We limit the scope of examining order effects to the neutral images we generated. For each neutral image, we count the consecutive positive (e.g. $P1$) and consecutive negative images (e.g. $N1$) which appear before it. If a neutral image appears before the current image then $P = 0$. A randomly sampled set of images most often produces sequences between $N5$ and $P5$ since the negative valence (e.g. SADNESS and ANGER) and positive valence (e.g. HAPPINESS) images are sampled with equal proportion in the random blocks of images.

Figure 4 shows the scatter plot of how annotator ratings are related to the number of previous positive or negative images seen. There is a positive correlation, Pearson correlation of 0.194 ($p < 1e^{-3}$), between the number of previous positive images seen and the label given. This suggests sequences of positive and negative images preceding neutral images, even ordered randomly, may prime annotators to select similar labels for these neutral images.

This positive correlation result could be influenced by the specific ordering of this particular block of images sampled and used for the experiment. For example, the "more positive" neutral images may follow longer sequences of positive images by chance. To address this potential short coming, we used five-fold shuffled versions of the images to check whether

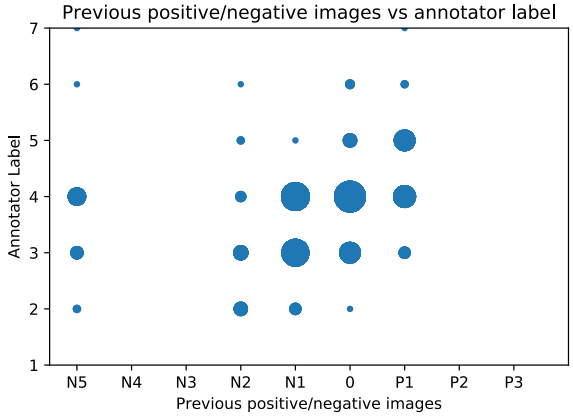

Fig. 4. Plot of the number of previous positive (P) and negative (N) images vs annotator ratings of neutral images (NO EMOTION) for the set A permutation A. The size of the markers represent frequency that the label was selected

this effect is consistent across all five folds. We also added a second set of randomly sampled images shuffled across five-folds. Table I shows the various correlation coefficients across the various permutations of the images across two sets of images. This table illustrates that varying the *set* of P (number of positive/negative images preceding the example) and the position of individual images has a large effect on the correlation results. In 40% of these randomly-sampled sets and randomly-shuffled permutations, there are statistically significant correlations between the number of previous negative or positive examples seen and the resulting label.

TABLE I
PEARSON CORRELATION 5 ORDER PERMUTATIONS FOR EACH SET OF RANDOMLY SAMPLED IMAGES

	Set A Images		**Set B Images**	
Permutation	**Pearson Corr.**	**P-Value**	**Pearson Corr.**	**P-Value**
a	**0.194**	0.001	0.030	0.764
b	**0.329**	0.020	**0.198**	0.030
c	**0.332**	0.019	-0.182	0.071
d	0.022	0.877	-0.057	0.577
e	0.170	0.194	-0.191	0.057

Examining the specific set of neutral images from which Table I was derived, the mean valence rating of these images ranged from $3.42 - 4.92$. This large variation in the perceived valence of neutral images suggests the need to compare this effect of preceding images on the same neutral image.

For each neutral image, we find the number of preceding negative and positive images across all folds of the experiment. Using the resulting label and the number of preceding positive images, we again compute the correlation between the two. However, due to the small sample set ($15 < n < 25$), the correlation values lacked statistical significance. In addition, some images only include negative preceding images or only positive preceding images due to the random shuffling of the image order in the original design of the experiment.

Experiment 3: Ordering effects in Controlled Set

To isolate the effect of varying lengths of positive and negative image sequences preceding the same neutral image, we use an alternative experimental design. In this design, each annotator is presented with 5 blocks of images to annotate. 2 of these 5 are random blocks in which an evenly mixed number of HAPPINESS, SADNESS, ANGER, and NO EMOTION poses are randomly shuffled. The other three blocks are pairs of alternating blocks in which version A displays k positive images preceding a neutral image while version B presents k negative images preceding a neutral image. Figure 5 illustrates the experimental set up. This configuration allows us to collect two sets of labels for the same neutral image: one set after an annotator has seen a negative sequence of images and one set after an annotator has seen a positive sequence of images. We also vary k, the number of negative/positive images preceding neutral images, from $k = 2$ to $k = 4$.

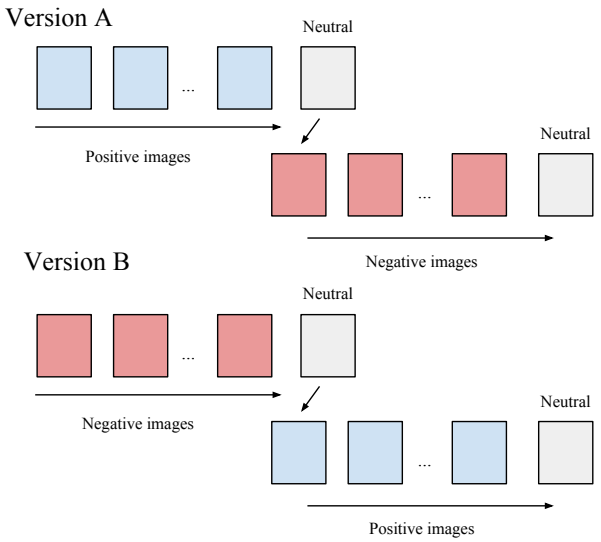

Fig. 5. Experimental design of controlled set of images. Version A presents k positive images before the first neutral image and k negative images before the second neutral image and so on. Version B prevents an inverted version.

1) Effect of previous image: We first examine whether a neutral image label is related to the previous image. The previous image label, should be similar to, but may be different from, the actual ground truth valence of the image. The ground truth valence is based on the parameter used to generate the image with StyleGAN (see Section III). Table II shows the linear correlation of image labels with the previous image label and the previous image ground truth valence. For all values of k, we observe a positive correlation with the previous label given. This again supports the hypothesis that decision-making in the context of emotion labeling can be susceptible to affective priming. The correlation with the actual ground truth valence is mostly positive but not statistically significant. This suggests that the current label an annotator provides is more correlated with the label they gave to the immediate preceding image (perceived valence) than with the ground truth valence.

TABLE II
PEARSON CORRELATION BETWEEN CURRENT LABEL AND I) PREVIOUS LABEL, II) PREVIOUS VALENCE USED TO GENERATE IMAGE. THE STATICALLY SIGNIFICANT VALUES ($P < 0.05$) ARE BOLDED

k-length	Corr. with prev. label	Corr. with prev. valence
$k = 2$	**0.115**	0.180
$k = 3$	0.063	-0.01
$k = 4$	**0.125**	0.116

2) Effect of previous sequence of positive and negative images: Collecting responses for both conditions in Figure 5, we can compare whether there are differences between the negative and positive conditioning on the labels for the same neutral image. For example we can compare the mean between the two conditions for image j: $Px_j = \frac{1}{n_{pos}} \sum_{i \in pos} x_{ij}$ for neutral image labels in the positive condition, and $Nx_j = \frac{1}{n_{neg}} \sum_{i \in neg} x_{ij}$ for the neutral image labels in the negative condition. Table III summarizes the percentage of positive, negative, and zero-valued difference images across different values of k. Here, most images exhibit a higher mean label value after the positive priming condition than the negative priming condition. This again supports the positive auto-correlation phenomenon from affective priming. Applying the t-test for difference in valence means across each image example with Bonferroni correction does not yield statistically significant results for any image. This could be due to a small sample size (i.e. 10 responses for each condition) and the limited number of values annotators could select for each image since we discretized valence labels (i.e. 1, 2, ... 7).

TABLE III
PERCENTAGE OF IMAGES FOR EACH k WHERE THE DIFFERENCE BETWEEN THE POSITIVE (Px_j) AND NEGATIVE (Nx_j) CONDITIONS ARE GREATER THAN, LESS THAN, AND EQUAL TO ZERO.

k-length	$Px_j - Nx_j < 0$	$Px_j - Nx_j = 0$	$Px_j - Nx_j > 0$
$k = 2$	30.95%	11.90 %	57.14%
$k = 3$	43.33%	6.67 %	50.00%
$k = 4$	37.50 %	4.17 %	58.33%

The images which generate negative differences between the Px and Nx conditions are fairly consistent across the different lengths of k. For example, the same neutral image will take a higher mean valence after negative images than after positive images across for $k = 2, 3, 4$. This suggests that there are features inherent to the displayed images that may obfuscate the positively auto-correlated effect observed in earlier experiments. Figure 6a illustrates an example where the label after a positive sequence (i.e. Px condition) of images is consistently more positive than the mean image label after a negative sequence (i.e. Nx condition) of images (i.e. affective priming effect). Figure 6b illustrates an example where the opposite is true. Here, contrast effects may dominate the labels in the two conditions; the 4^{th} image in the first row looks more negative compared to the first three images in the first row than

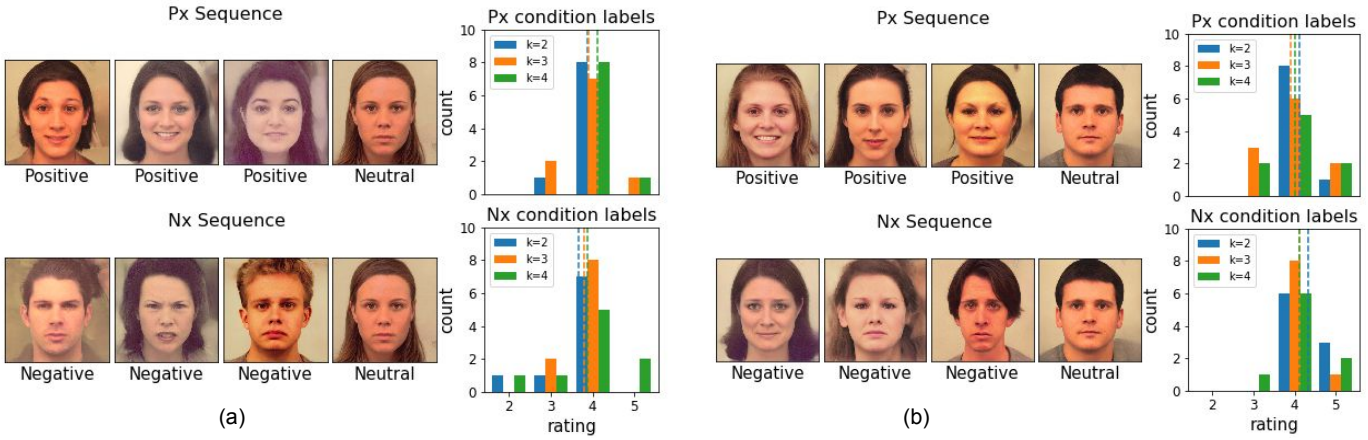

Fig. 6. Example of Px and Nx sequences ($k = 3$): a) a neutral image which consistently generates higher valence labels after sequences of positive images than after sequences of negative images. b) neutral image generating higher valence after sequences of negative images than sequences of positive images

compared to the first three images in the second row.

V. Discussion

Q1: Inter-annotator agreement

Comparing entropy and Fleiss' kappa, a set of shuffled sequences of examples seen by annotators yields lower inter-annotator agreement than a uniformly ordered sequence of examples. This result confirms and complements recent work finding that annotators produce different labels for examples in context than in randomized examples [26].

Q2: Example ordering in randomly shuffled sequences

To investigate systematic biases in labels due to example ordering, we tested correlation between the number of positive/negative images appearing before a neutral example and the label an annotator gives to the neutral example. We find significant correlation, supporting the hypothesis that affect priming from sequences of negative/positive valence images produces positive auto-correlation in examples. In large scale data labeling tasks, these sequences of images that appear due to random shuffling of examples may inadvertently introduce bias into the image labels. We test the robustness of this finding by shuffling two sets of images so that the same image follows sequences varying in length and valence. We observe that auto-correlation of an image set can vary drastically across different orderings. Since only limited few orderings are shown to annotators in a typical labeling task, the resultant labels can contain bias due to unintentional affective priming effects.

Q3: Effect of previous images on current image label

In our controlled experiment, annotators are assigned to randomized conditions of positive/negative image sequences preceding neutral images. We find significant correlation between the current label of neutral images and the previous label given. The mean label for a neutral image is mostly higher following positive sequences than negative sequences. However, there is a portion of images that consistently exhibit the opposite effect. This result suggests that only looking at the labels is not sufficient for completely characterizing sequential effects. Qualities inherent to a specific image (i.e. emotion ambiguity, facial structure) may influence the resulting label more than any ordering effects. While we find evidence to support affective priming in sequential labeling of facial images, there may also be other potent effects such as instance-dependent assimilation and contrast effects [14].

VI. Conclusion

In practice, long sequences of images are frequently randomized and given to very few annotators for annotation. In this work, we find that some orderings of randomly shuffled sequences of images may significantly bias annotator labels. We find evidence to support positive auto-correlation between labels; an effect consistent with affect priming rather than the Gambler's fallacy. This effect is important to consider when reducing systematic label noise in subjective labeling tasks. Thus, randomizing sequences of images for labeling may be inadequate to remove annotator bias.

Future data collection in affective computing tasks would benefit from collecting both annotation order and annotator information. Assigning multiple annotators to each of several orderings of examples would better allow analysis of sequential effects than randomizing orderings across all annotators. Furthermore, intentionally ordering examples to mitigate affective priming may reduce the risk of systematic biases in labels and limit the source of label noise to inherent example ambiguity. Future work to understand sequential effects in affective computing should include features of specific examples which could better encapsulate instance-dependent noise. Stratified sampling according to characteristics such as race, gender, and facial features would help further examine the presence of affect priming and contrast effects.

References

[1] H. J. Jung and M. Lease, "Improving quality of crowdsourced labels via probabilistic matrix factorization," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[2] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2008, pp. 254–263.

[3] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research*, vol. 11, no. Apr, pp. 1297–1322, 2010.

[4] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 28, no. 1, pp. 20–28, 1979.

[5] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, no. 1, pp. 1–1.

[6] A. Mollahosseini, B. Hasani, M. J. Salvador, H. Abdollahi, D. Chan, and M. H. Mahoor, "Facial expression recognition from world wild web," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 58–65.

[7] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5562–5570.

[8] C. Frings and D. Wentura, "Trial-by-trial effects in the affective priming paradigm," *Acta Psychologica*, vol. 128, no. 2, pp. 318–323, 2008.

[9] D. L. Chen, T. J. Moskowitz, and K. Shue, "Decision making under the gamblers fallacy: Evidence from asylum judges, loan officers, and baseball umpires," *The Quarterly Journal of Economics*, vol. 131, no. 3, pp. 1181–1242, 2016.

[10] R. Morris, D. McDuff, and R. Calvo, "Crowdsourcing techniques for affective computing," in *The Oxford handbook of affective computing*. Oxford Univ. Press, 2014, pp. 384–394.

[11] I. Siegert, R. Böck, and A. Wendemuth, "Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements," *Journal on Multimodal User Interfaces*, vol. 8, no. 1, pp. 17–28, 2014.

[12] S. Parthasarathy and C. Busso, "Preference-learning with qualitative agreement for sentence level emotional annotations," *Proc. Interspeech 2018*, pp. 252–256, 2018.

[13] R. Lotfian and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 415–420.

[14] S.-M. Hsu and L.-X. Yang, "Sequential effects in facial expression categorization." *Emotion*, vol. 13, no. 3, p. 573, 2013.

[15] H. J. Jung, Y. Park, and M. Lease, "Predicting next label quality: A time-series model of crowdwork," in *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

[16] P. Donmez, J. Carbonell, and J. Schneider, "A probabilistic framework to learn from multiple annotators with time-varying accuracy," in *Proceedings of the 2010 SIAM International Conference on Data Mining*. SIAM, 2010, pp. 826–837.

[17] M. Atcheson, V. Sethu, and J. Epps, "Demonstrating and modelling systematic time-varying annotator disagreement in continuous emotion annotation," *Proc. Interspeech 2018*, pp. 3668–3672, 2018.

[18] I. Siegert, R. Böck, and A. Wendemuth, "The influence of context knowledge for multi-modal affective annotation," in *International Conference on Human-Computer Interaction*. Springer, 2013, pp. 381–390.

[19] W. Garner, "An informational analysis of absolute judgments of loudness." *Journal of experimental psychology*, vol. 46, no. 5, p. 373, 1953.

[20] D. A. Leopold, G. Rhodes, K.-M. Müller, and L. Jeffery, "The dynamics of visual adaptation to faces," *Proceedings of the Royal Society B: Biological Sciences*, vol. 272, no. 1566, pp. 897–904, 2005.

[21] S. Chang, C.-Y. Kim, and Y. S. Cho, "Sequential effects in preference decision: Prior preference assimilates current preference," *PloS one*, vol. 12, no. 8, p. e0182442, 2017.

[22] A. L. Bouhuys, G. M. Bloem, and T. G. Groothuis, "Induction of depressed and elated mood by music influences the perception of facial emotional expressions in healthy subjects," *Journal of affective disorders*, vol. 33, no. 4, pp. 215–226, 1995.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[24] E. Goeleven, R. De Raedt, L. Leyman, and B. Verschuere, "The karolinska directed emotional faces: a validation study," *Cognition and emotion*, vol. 22, no. 6, pp. 1094–1118, 2008.

[25] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *arXiv preprint arXiv:1812.04948*, 2018.

[26] M. Jaiswal, Z. Aldeneh, C.-P. Bara, Y. Luo, M. Burzo, R. Mihalcea, and E. M. Provost, "Muse-ing on the impact of utterance ordering on crowdsourced emotion annotations," *arXiv preprint arXiv:1903.11672*, 2019.