Can the Crowd Tell How I Feel?
Trait Empathy and Ethnic Background in a Visual Pain Judgment Task

Maria Matsangidou
University of Kent
Kent, UK
M.Matsangidou@kent.ac.uk

Jahna Otterbacher
Open University of Cyprus
Nicosia, Cyprus
Jahna.otterbacher@ouc.ac.cy

Chee Siang Ang
University of Kent
Kent, UK
C.S.Ang@kent.ac.uk

Panayiotis Zaphiris
Cyprus University of Technology
Limassol, Cyprus
panayiotis.zaphiris@cut.ac.cy

ABSTRACT Many advocate for artificial agents to be empathic. Crowdsourcing could help, by facilitating human-in-the-loop approaches and dataset creation for visual emotion recognition algorithms. Although crowdsourcing has been employed successfully for a range of tasks, it is not clear how effective crowdsourcing is when the task involves subjective rating of emotions. We examined relationships between demographics, empathy and ethnic identity in pain emotion recognition tasks. Amazon MTurkers viewed images of strangers in painful settings, and tagged subjects' emotions. They rated their level of pain arousal and confidence in their responses, and completed tests to gauge trait empathy and ethnic identity. We found that Caucasian participants were less confident than others, even when viewing other Caucasians in pain. Gender correlated to word choices for describing images, though not to pain arousal or confidence. The results underscore the need for verified information on crowdworkers, to harness diversity effectively for metadata generation tasks.

Keywords. Crowdsourcing · Ethnicity · Pain · Distress · Empathy · Image metadata

1 Introduction

Many advocate for artificial agents and systems to be more empathic in their interactions with humans. Machines that can recognize emotions stand to play a significant role in the development of next-generation human-computer interaction systems [10; 14; 34; 63]. Furthermore, with the emergence of social media, users are uploading millions of pictures everyday trying to express their emotions and thoughts with others. For example, Whatsapp users are uploading 700 million new photos per day, Facebook's users share 350 million new photos each day while Snapchat emerges on top, with total share of 8,796 photos per second [49]. This illustrates clearly the need to consider the quality of metadata generation. Whether the intention is to develop algorithms that infer image properties, or to use human-in-the-loop approaches (i.e., relying on the perceptual abilities of online crowdworkers in real-time [31], in images depicting people, ensuring metadata quality is crucial. In the current work, we consider a metadata generation task that is more challenging than the labeling of image content (i.e., whether the image depicts a person, animal or object). Specifically, we are interested in emotion recognition from images of people in pain and distress.

It must be noted that automated emotion recognition, which contributes to many high-stake applications involving behavioral analysis in both the commercial and medical domains, is not without controversy, given the potential risks. Imagine a robot designed to offer communication support to an individual with depression, which embeds emotion recognition technology. In such a context, the costs of misrecognition of the user's distress are very high, with many potential consequences. For example, if the robot was to misrecognize a sentiment such as disgust for sadness or hostility, this could lead to inappropriate responses on behalf of the agent, such as offending a patient who may already be in a sensitive state.

In most settings, computer-based emotion recognition is achieved in one or two ways. Emotions can be detected indirectly through observing the other party's facial expressions, gestures and voice (including verbal and written communication) [57], or directly from physiological signals such as heart rate [51]. In the current study, the task is to recognize the negative emotions of an individual depicted in an image, where the annotator (or the "crowd") must rely on visual cues only, in performing indirect emotional distress recognition.

We consider the possibility of using crowdsourcing for emotional distress recognition on a mass scale, as facilitated by a popular crowdsourcing platform (Amazon Mechanical Turk), to label images of people with respect to what they are feeling. Several studies have been conducted on emotion recognition, and found that people tend to be relatively accurate at judging facial expressions [15; 57]. However, to the best of our knowledge, only few studies have explored the detection of emotions based on online pictures/videos, with no additional cues. Achieving a better understanding of the task and the characteristics of individuals who can perform it reliably and accurately stands to benefit human-computer interaction systems and their ability to become more empathic.

1.1 Crowdsourcing Emotion Judgments

Crowdsourcing has emerged as an effective means to complete small, well-defined tasks, which while simple for humans, cannot yet be performed reliably by machines. Essentially an open call for labor with flexible contractual arrangements [47], crowdsourcing provides a convenient, low-cost solution for obtaining specific feedback that is arguably objective and valid [38; 69]. Crowdsourcing is already being used to generate annotations for images depicting human behavior with reported success [27; 53; 54].

Crowdsourcing also has the potential to enable us to gather information that is universal and relevant across cultures [43], given the ease in reaching diverse groups of crowdworkers via popular platforms. Nonetheless, the diversity of "the crowd" is precisely the characteristic that may challenge the task of accurately identifying a person in distress and/or pain. As will be explained, psychological studies show clearly that who we are (i.e., our demographic characteristics and personalities) relate to our abilities to understand others' emotions. In addition, we typically have more affinity towards, and an enhanced ability to understand, those more like ourselves. Given

the diversity of the crowd, how can we ensure that high-quality metadata on emotional distress labeling tasks are generated?

1.2 Goals of the Current Work

This study gauges the feasibility of crowdsourcing on a visual pain recognition task (i.e., images) and an ethnically diverse workforce. Previous research has explored emotion recognition from speech and textual analysis [25; 33]. Since visual recognition has not been extensively explored in the context of crowdsourcing, our goal is to see how factors such as demographics, personality traits (i.e., level of empathy), and the degree to which one identifies with his or her ethnic group impacts one's approach and performance on the visual pain recognition task. In short, we address four novel research questions:

- Can "crowdworkers" recognize depicted people in pain and distress? How do demographics and ethnicity of the worker and the target individual depicted in the images impact performance?
- How does the empathy level of the worker impact performance on task?
- How does the strength of one's identity to his/her ethnic group impact performance?

The study aims to achieve the necessary understanding of the relationship between the task of image emotion recognition, the social cues surrounding the task (i.e., how in-group or out-group status affects the empathic response of the annotator) and the quality of image metadata that we might expect to derive via crowdsourcing.

2 Literature Review & Hypotheses

The psychological construct of empathy refers to the ability to understand and share positive or negative emotions. It is a developmental emotion that first appears in infancy, and promotes pro-social behaviors [70]. Levenson and Ruef [37] outline three key components of empathy: "(a) knowing what another person is feeling, (b) feeling what another person is feeling, and (c) responding compassionately to another person's distress".

The success of the emotion recognition task primarily hinges on the first of the above three components; it is clear that recognizing the emotions of another requires empathy. However, most of what we know about empathic responses to others concerns face-to-face communication, where around 90% of emotional expressions are communicated non-verbally [20]. In this setting, empathic responses are the result of careful observation of both the verbal and non-verbal cues during communications with the other (i.e., target individual).

In our task, annotators (participants) are asked to infer a stranger's negative emotion, based only upon visual cues available in a still image. Given the close connection between empathy and pain emotion recognition, we expect that individuals who have greater levels of trait empathy (i.e., a more empathic personality), will be better

annotators as compared to those with lower levels of empathy. We also expect better annotation results when annotators and depicted individuals are of the same ethnic group. Obviously, what constitutes a "quality annotation" depends on the intended use of the image metadata. Therefore, we consider multiple response variables.

As explained in the methodology section, we consider one measure of participants' emotional reaction to a painful image (their reported level of pain arousal) as well as a measure of their beliefs about their ability to describe the subject's emotion accurately (confidence). Next, we consider the affective content of the tags participants use to describe subjects' emotions. Following Warriner and colleagues [67], we consider the valence (i.e., pleasantness) of word tags, their degree of arousal (i.e., the intensity of emotion expressed in a chosen tag), as well as word dominance (i.e., the degree of control suggested by a chosen word). In other words, the current work explores these five response variables (pain arousal, task accuracy, as well as affective meanings consisting of valence, arousal and dominance). In the remainder of this section, we describe the background and motivation of our work as well as the hypotheses to be tested.

2.1 Empathy and reaction to others' suffering

Empathy manifests itself as a reaction to others' emotions. We examine the empathic responses to others' suffering, and more specifically, responses to the primary emotion of sadness conveyed through images of pain and distress. Sadness is a fundamental emotion experienced by all human beings [12; 15], and is most strongly associated with the understanding of a permanent loss. A particularly good example is death, which has the ability to transform the individual's interpretation of life and world [60]. The importance of the phenomenon is highlighted by the simultaneous experience of several emotions regarding the individual's shattering: anxiety, irritation-anger, emptiness, worthlessness, meaninglessness, hopelessness, weakness, brokenness and/or guilt [5; 21; 46].

However, while experienced by everyone, recognizing sadness or pain in another is not necessarily easy. The perception of pain is based on "representations" that one has made. Each individual creates and stores mental representations of personal pain experiences. These representations are later called upon, to identify the perception of pain expressed by others [3; 28]. For these reasons, we expect to find that who someone is, correlates to his or her ability to infer the target person's emotion.

H1a: Demographic characteristics such as gender and age correlate to participants' level of pain arousal.

H1b: Demographic characteristics such as gender and age are correlated to confidence on task.

H1c: Demographic characteristics of gender and age are correlated to the affective content of words used to describe painful images.

Women generally self-report as being more emotional than men do [61] and are more empathic than men [35]. Of particular note is that the experience of negative

emotions, such as sadness and pain, are most often reported by women [7; 18; 26], and the duration of the feeling seems to be longer in women than men [53]. Finally, women express and interpret emotions more accurately [22; 23; 42], and girls express their sadness more intensely than do boys [11].

In contrast to gender, few studies have investigated the correlation between emotions, empathic reactions and age. Some findings support the notion that empathy is a pro-social characteristic that appears and develops throughout life. Specifically, psychologists consider the mechanism of crying in infants as an empathic reaction, with female infants empathizing more than males [40]. In addition, empathy may decrease as one approaches adulthood [55], then increase again in old age, with older people exhibiting higher scores on standardized measures of empathy [39; 56]. We expect that individuals with higher levels of trait empathy will be better at our visual pain recognition task as compared to those with lower empathy. We also expect to find that such individuals will describe image subjects' emotions more intensely.

H2a: Empathic individuals will experience greater arousal as compared to less empathic individuals, when viewing images of others in pain.

H2b: Empathic individuals will report greater confidence on task as compared to less empathic individuals.

H2c: Empathic individuals will describe images of others in pain using more intense words as compared to less empathic individuals.

2.2 Empathic reaction to out-group members' feelings

Facial expressions of emotions fall into two basic categories: universal and culturally specific. As a consequence, the origin and ethnicity of the individual can affect his or her non-verbal communication behavior. We should not expect that a diverse group of individuals would express a given emotion in the same manner. Indeed, ethnic or cultural differences, defined as individuals having been positioned within a given in-group while being excluded from one or more out-groups, have been shown to correlate to emotional behavior [64].

Likewise, research suggests that individuals' abilities to accurately recognize another's emotions relates to group membership and similarity. Increased similarity as well as identification with others can lead to increased sharing of the experience and hence heightened empathy [52]. In particular, belonging to a social group serves as a form of contingent that enhances empathy among group members [6; 24; 65]. By including the other group members as part of one's self-concept [2], people are generally able to empathize more strongly with in-group members. This phenomenon is called in-group advantage [16; 17].

The in-group advantage can lead to increased accuracy in visual emotion detection for members of the same ethnicity [66; 68]. Similar results were obtained with regards to emotional detection of speech. Specifically, individuals within the same country but of a different culture or ethnicity (e.g., White Canadians and Canadian Aboriginals) could not detect one another's emotions as accurately as those within the same group [1].

Finally, based on studies of people's reactions to others as depicted through images, it is evident that empathic responses toward those suffering were stronger within in-group members and weaker for out-group members' suffering [8; 19]. As a result, individuals' responses when viewing images of in-group members were more empathic than when they viewed images of those from an out-group [4].

H3a: Participants report greater pain arousal when viewing images of in-group (versus out-group) members.

H3b: Participants report greater confidence when describing emotions of in-group (versus out-group) members.

H3c: Participants will describe the pain of in-group members using more intense word labels, as compared to members of their out-group.

3 Data & Method

Our visual pain recognition task consisted of three parts. After viewing an image depicting a stranger in pain or distress, participants (1) rated their own level of pain arousal, (2) described the emotional content of the image via open-ended tagging, and (3) assessed their performance on the tagging task. Participants also completed a questionnaire concerning their demographic background, as well as two standard psychological questionnaires: Davis' Interpersonal Reactivity Index (IRI) and the Multigroup Ethnic Identity Measure (MEIM). Relevant details are provided in the following sub-sections.

3.1 Image Dataset

We used a dataset of 42 images developed by a team of neuroscientists, who used them to investigate the neural basis of reactions to depicted subjects [41]. The images depicted East Asian (EA), African (AA) or Caucasian (CA) American subjects. Thirty-six images were painful (e.g., a woman crying during a flood) and six were neutral (e.g., a man enjoying an outdoor picnic) situations. We include neutral images to permit participants' arousal level to "settle down", (e.g., to avoid habituation effect). In a previous experiment using this dataset, participants were asked to indicate "how badly" they feel for the main subject(s) in the image on a 4-point scale. The results showed that the images elicit both reliable and valid responses. Example images are shown in Figure 1.

3.2 Participants

We used Amazon Mechanical Turk (MTurk) to recruit a crowdsourced workforce. We selected MTurk since it has been successfully used to crowdsource annotations on text, scenes, pictures [9; 27; 62] and emotions [46]. We targeted three groups of participants by their self-reported ethnic background (EA, AA or CA).

Participants had to be native or near-native English speakers who reside in the United States. They were rewarded with $5 for their time, and all but one took less than 60 minutes. A total of 120 participants, ranging in age from 18 to 57 years (M_{age} = 29.88, SD = 7.55) completed the study. Specifically, 30 East Asian-American (EA) (Males = 13, Females = 17), 39 African-American (AA) (Males = 22, Females = 17), and 51 Caucasian-American (CA) (Males = 31, Females = 20), participated.

3.3 Experimental Design, Tasks and Validation

The 42 images were shown to each participant, in the same random order. After viewing an image, participants completed the pain arousal item ("How badly do you feel for the person(s) in the image?"). Next, they were asked to provide three emotion tags ("How would you describe the emotions of the main subject(s) of the image?"). Finally, participants were asked to rate their confidence level ("How confident are you that you accurately described the emotion(s) of the main subject in the image?"). The first and third tasks used a four point Likert item (1=not at all to 4=very much).

To confirm the validity of our approach, factor analysis was used to examine the structure of the image characteristics, as the images are being used to stimulate a response in participants, along with pain arousal scores. The analysis revealed a solution that explained 59.426% of variance and that had structural coefficients (loadings) > .50 for all factors. Varimax rotation yielded three factors, corresponding to the ethnicity of the subjects (EA, AA, and CA pain arousal pictures), and consisting of 12 items each. This analysis also revealed a high degree of reliability and validity. The internal consistency of each item measured by Cronbach alpha, the EA Pain Arousal α was .932 with an eigenvalue of 10.726, the AA Pain Arousal α was .915 with an eigenvalue of 9.695, and the CA Pain Arousal α was .926 with an eigenvalue of 4.538.

Fig. 1. Example images of EA, AA, and CA individuals in painful settings

In addition, the reliability and validity of the image characteristics were re-tested with the self-reported confidence scores. This yielded a solution that explained 48.736% of variance and that had structural coefficients > .50. Varimax rotation again yielded three factors in this case, based on the ethnicity of the subjects (EA, AA, and CA) and consisting of 14 items each. This analysis also revealed a high degree of both reliability and validity. In particular, the internal consistency of each item measured by Cronbach alpha, the EA Task Confidence α was .894 with an eigenvalue of 7.783, the AA Task Confidence α was .896 with an eigenvalue of 6.701, and the CA Task Confidence α was .866 with an eigenvalue of 5.986.

3.4 Psychological Tests

Davis' Interpersonal Reactivity Index (IRI).

Davis' IRI [13] is a measure of dispositional (or trait) empathy that considers a set of four distinct, though related constructs. Each of its four subscales (empathic concern, fantasy, perspective taking and personal distress) was assessed with 7 items on a 5-point Likert scale (0 = does not describe me well to 4 = describes me very well). The subscales that pertain to cognitive dimensions of empathy, the fantasy subscale (FS) and the perspective taking (PT) subscale, measure the tendency to get caught up in fictional stories and imagine oneself in the same situations as these fictional characters, and the tendency to take the psychological point of view of others, respectively. The empathic concern and personal distress subscales measure the affective dimensions of empathy. Specifically, the empathic concern (EC) scale measures sympathy and concern for others and is typically considered an other-oriented emotional response in which attention is directed to the person in distress [59]. On the contrary, the Personal Distress (PD) scale is considered a self-oriented emotional response in which attention is directed at one's negative emotions of distress and the reduction of these negative emotions.

Others have found the IRI instrument to have a high degree of reliability and validity, which was also supported by our findings. We used exploratory factor analysis to examine its structure. This yielded a solution that explained 62.724% of variance and that had structural coefficients > .50 for all factors. Varimax rotation yielded four factors (EC, PD, PT and FS), consisting of seven items each. Furthermore the analysis revealed a high degree of both reliability and validity. Notably, the internal consistency of each item measured by Cronbach alpha, the Empathic Concern (EC) α was .917 with an eigenvalue of 5.030, the Personal Distress (PD) α was .888 with an eigenvalue of 4.426, the Perspective Taking (PT) α was .882 with an eigenvalue of 4.157, and the Fantasy Scale (FS) α was .853 with an eigenvalue of 3.949.

Multigroup Ethnic Identity Measure (MEIM).

The Multigroup Ethnic Identity Measure (MEIM) [50] is an instrument that reveals a high degree of reliability and validity in measuring the feelings and reactions of the individual, in relation to his or her reported ethnic group. The instrument contains questions designed to assess two related constructs. Participants answered 12 closed response items on a 4-point Likert scale (1 = strongly disagree to 4 = strongly agree). The first construct is relevant to Affirmation, Belonging and Commitment, gauges knowledge of and feelings toward one's ethnic group and consists of seven items (e.g., "I feel good about my cultural or ethnic background"). The second construct is relevant to Ethnic Identity Search and consists of five questions (e.g., "I think a lot about how my life will be affected by my ethnic group membership"). There are also two categorical items, where one is asked to select his or her ethnic group and that of his or her parents. Finally, in one open question, the participant is asked to state his or her ethnic group ("I consider myself to be…").

The high degree of reliability and validity of the instrument was supported. Factor analysis was used to examine the structure of the MEIM questionnaire. This yielded a

solution that explained 66.471% of variance and that had structural coefficients > .60 for all factors. Varimax rotation yielded two factors (Ethnic –Identity Search and Affirmation – Belonging – Commitment), consisting of five and seven items respectively. Furthermore, the analysis revealed a high degree of reliability and validity. Particularly, the internal consistency of each item measured by Cronbach alpha, the Ethnic –Identity Search α was .850 with an eigenvalue of 3.734, and the Affirmation – Belonging – Commitment α was .914 with an eigenvalue of 4.243.

3.5 Affective Content of Emotion Tags

We collected a total of 12,960 word tags (i.e., three tags for 36 painful images for 120 participants), which described the emotions of the main subject(s) of each image. Given the size of the corpus, we needed a means to automatically analyze the affective content expressed through the tags. It can be noted that *sentiment analysis*, or the detection of affect in textual communication, is a very active area of research in recent years, particularly among information retrieval [e.g., 37; 49] and natural language processing [e.g., 45] scholars. To this end, many resources, including sentiment lexicons, have been developed, in order to enable the exploitation of the rich sources of textual data shared via social media. However, as we aimed to examine the affective content of individual word-tags, and their correlation to participants' demographics and personal characteristics, we selected a lexicon developed by a team of psycholinguists, which aims to depict the affective norms of individual words [67], which is close in spirit to our task.

This resource is a collection of ratings on three affective dimensions for nearly 14,000 English words. As mentioned, the dimensions are valence, arousal and dominance. In Warriner et al. [67], participants rated a given word on a scale of 1 to 9, reflecting their feelings when reading the word, as follows:

Valence: How happy / pleased / satisfied / contented / hopeful do you feel?

Arousal: How excited / stimulated / frenzied / jittery / wide-awake / aroused do you feel?

Dominance: How controlled / influenced / cared-for / awed / submissive / guided do you feel?

Table 1 provides examples of words that score relatively high and low on each of the three dimensions. Specifically, what is shown is the mean score assigned by all participants in the study of Warriner et al. [67] who rated the given word.

For each word that our participants used as an emotion tag, we obtained the three affective scores in order to explore how personality and background might influence the words someone uses to describe another in pain. In total, 83% of our tags were found in the lexicon and have valid scores, leaving us with 10,704 word tags to analyze.

Table 1. Example words and their mean scores on three affective dimensions.

	High	Low
Valence	Happiness (7.05)	Disaster (2.97)
	Joyful (7.05)	Mourning (3.64)
Arousal	Thrill (7.19)	Calm (1.67)
	Panicky (7.00)	Dull (1.67)
Dominance	Strength (7.42)	Defeated (2.43)
	Courageous (7.38)	Rejected (2.43)

3.6 Statistical Analysis

We used parametric analyses (including correlation (Pearson's r), t-tests, one-way ANOVA and linear regression) to explore the relationships between participants' demographic characteristics, their levels of empathy and their ethnic identities and each of the five response variables, i.e. (1) pain arousal ratings, (2) self-assessed task confidence, and the (3) valence (4) arousal and (5) dominance of tags used to describe the emotions of strangers depicted in painful settings.

As the first two response variables were left-skewed, we applied the following transformation before performing our analyses: $(x + 1)^2$. In contrast, the scores on the three affective dimensions of word tags are right-skewed and thus were transformed as follows: $\log(x + 1)$.

4 Results

4.1 Demographic Characteristics of Image Annotators

Our first set of hypotheses (H1) proposed that annotators' demographic characteristics, and in particular, age and gender, are correlated to their performance on the visual pain recognition task. The literature suggests that empathy levels vary with both gender and age; thus, one's ability to understand others' emotional pain should, in theory, correlate to his or her ability to recognize strangers' pain and distress. An independent two-group t-test reveals significant gender differences in IRI scores, with respect to emotional concern ($t = -2.265$, $p<.05$) and personal distress ($t = -4.787$, $p<.001$). Women's scores reveal them to be more in touch with others' feelings (EC), yet also more focused on their own negative feelings of distress (PD), as compared to men. However, correlation analysis showed no significant correlation between any of the IRI scores and age.

Given these findings, we again used the t-test to compare each response variable across gender. As shown in Table 2, which details the mean/median scores by gender, we find no significant gender differences for pain arousal reported by participants, or for their self-reported confidence on task. We do, however, find differences on two of the affective dimensions of word tags assigned to images. Women tend to use words suggesting more arousal or excitement (e.g., panicky, dangerous, tragedy, rage) as

Table 2. T-tests comparing 5 response variables by gender. (****p<.001; **p<.01*)

	Men	Women	t
Pain	106.1/110.5	113.4/114.5	-1.623
Accuracy	131.3/131	130.6/130.5	0.262
Valence	3.151/2.790	3.123/2.670	1.195
Arousal	4.605/4.640	4.678/4.715	-3.384***
Dominance	4.163/3.850	4.116/3.840	2.642**

compared to men, whose chosen tags tended to rate higher on dominance (e.g., strength, courageous, understanding).

There were no significant correlations between participant age and either pain arousal scores, self-reported confidence, tag valence or dominance score. There was a statistically significant, albeit weak correlation between participant age and tag arousal score ($r = 0.0192$, $p<.05$).

The analysis leads us to reject hypotheses H1a and H1b concerning gender and age. In contrast, we do observe that gender is correlated to the types of words (i.e., tags) that participants choose to describe the emotions of the depicted subjects in painful images. Thus, we support H1c.

4.2 Empathy Levels of Image Annotators

We hypothesized that more empathic individuals will be better annotators in the emotion detection task, because they have an easier time knowing and feeling what another feels (H2). Table 3 details a linear regression analysis in which each of the five response variables was regressed on the four IRI scores as well as participant gender. It is clear that emotional concern (EC), the dimension of empathy that reflects one's ability to understand another's feelings, is positively related to our first two response variables, the pain arousal score and the self-reported task confidence. It is notable that EC plays a key role in explaining the variance of both pain arousal and task confidence, even when we control for gender, which we found to be highly correlated to EC and PD.

We also observe evidence of significant, albeit very weak correlations between EC and PD on the affective properties of word tags. However, the explanatory power of these models is almost nil. Our results support hypotheses H2a and H2b; participants who are other-oriented experience greater pain arousal when viewing images of strangers in pain, and report higher confidence in describing the strangers' emotions. We reject hypothesis H2c, since IRI scores explain almost zero of the variance in the valence, arousal and dominance scores of the word tags used to describe images.

Table 3. Linear regression model: response variables regressed on IRI scores. (****p<.001; **p<.01; *<.05*)

	EC	PD	FS	PT	Gender	F	R2
Pain	368.84***	35.09	-32.32	-16.59	117.4	9.386***	0.2916
Confidence	264.39**	32.46	-23.40	9.945	-1102.5	2.920*	0.1135
Valence	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Arousal	-0.00547*	0.00364*	0.00203	-0.000806	0.0650**	5.841***	0.00272
Dominance	-0.00477*	-0.0010	0.00208	0.004062*	-0.0263	3.011***	0.00141

Table 4. Mean / median responses by subject and participant ethnicity group.

	Participant Ethnicity								
	EA			**AA**			**CA**		
Subject Ethnicity	**EA**	**AA**	**CA**	**EA**	**AA**	**CA**	**EA**	**AA**	**CA**
Pain	3.13/3.3	3.07/3.2	2.83/2.8	3.25/3.3	3.29/3.4	3.05/3.1	2.96/3.1	2.98/3.1	2.86/3
Confidence	3.27/3.3	3.22/3.3	3.14/3.1	3.27/3.2	3.34/3.3	3.21/3.2	2.91/2.9	3.02/3.0	2.93/2.9
Valence	2.96/2.5	3.12/2.6	3.26/2.7	3.00/2.6	3.20/2.8	3.21/2.8	3.05/2.8	3.19/2.8	3.21/2.8
Arousal	4.58/4.5	4.61/4.5	4.60/4.5	4.68/4.8	4.64/4.7	4.65/4.7	4.64/4.7	4.67/4.7	4.63/4.6
Dominance	3.99/3.8	4.08/3.8	4.20/3.8	4.08/3.8	4.15/3.8	4.23/3.9	4.11/3.9	4.17/3.9	4.20/3.9
Pain									
Confidence		3.00/3.1			3.20/3.3			2.93/3.0	
Valence		3.74/3.8			3.81/3.8			3.44/3.4	
Arousal		3.12/2.6			3.14/2.7			3.15/2.8	
Dominance		4.60/4.5			4.66/4.7			4.65/4.6	
		4.09/3.8			4.16/3.8			4.16/3.9	

4.3 Reacting to Emotions of In- vs. Out-group Members

Having examined the correlations between annotator demographics and levels of trait empathy and our five response variables, we now consider the possible role of ethnic group and the greater in-group sensitivity. First, we can ask whether in general, there are differences in our five response variables with respect to participant ethnic background. Table 4 details the mean / median responses on each response variable, broken out by participant and image subject ethnicity. The last row of the table shows the

average responses by participant ethnicity only (i.e., collapsing the three categories of image subjects).

Considering only participant ethnicity, one-way ANOVA reveals no significant differences with respect to pain arousal scores, however, the self-reported confidence scores differ ($F = 5.817$, $p<.05$). Specifically, Tukey HSD reveals that both EA and AA participants report higher confidence as compared to CA ($p<.05$ for both). For the three affective dimensions of word-tags, there are significant differences only with respect to dominance ($F = 5.99$, $p<.05$). Here, we find that EA participants use words with lower dominance scores, as compared to either AA or CA participants ($p<.05$ for both).

Next, we consider the possible effect of the ethnicity of the subject depicted in pain. We divided the images into three groups according to subject ethnicity, as shown in Table 5. We then performed one-way ANOVAs separately on each ethnic group (EA, AA, CA) for each of the response variables, with participant ethnicity as the grouping variable. In the case of a significant ANOVA, Tukey HSD was used to determine which participant ethnic groups reacted differently to the set of images.

As shown, the ethnicity of both image subject and participant play a role in their performance on the visual pain recognition task. For example, with respect to images of African Americans (second column in Table 5) in painful settings, AA and CA participants experienced differing levels of pain arousal, as well as self-reported confidence on task. Table 4 confirms that Caucasian participants experienced less pain arousal, and report reduced confidence, as compared to African Americans.

Having observed that image subject and participant ethnicity are important factors in the visual pain recognition task, we move on to consider greater in-group sensitivity. Specifically, we use regression analysis, applied to the three sets of images (broken out by image subject ethnicity), as described above. We created three indicator variables (AA, EA and CA Participant) in order to model cases in which out-group participants are viewing a set of images depicting subjects from a different ethnic group. In addition, we examine whether the strength of the participant's ethnic identity (i.e., MEIM scores) mitigates the greater in-group sensitivity.

Table 9, Table 10 and Table 11 detail the regression models for images of EA, AA and CA subjects, respectively. With respect to participants' levels of pain arousal and their confidence on task, it is clear that the strength of their ethnic identity (MEIM-ID-search) is highly correlated to the response variables, more so than in-group / out-group status with respect to the subject of the image. Note that for AA and CA images, none of the models predicting the affective dimensions of word-tags were significant and are therefore not detailed.

We removed the MEIM variables from the regressions in order to see if the out-group member indicator variables would play a more significant role in explaining the variance in the response variables. These results are shown in Tables 6, 7 and 8. Here, we can see that CA participants tend to be less confident on task when viewing images of out-group members (i.e., in Tables 6 and 7, we observe negative, highly significant coefficients on the CA indicator variable) as compared to the respective in-group participants. Finally, both EA and AA participants report more confidence when describing the pain of CA subjects, as compared to the in-group (CA) participants. None

of the models concerning the affective content of word tags were significant and are therefore not shown.

Table 5. Significant group differences per post-hoc Tukey HSD. *(**p<.01; *p<.05)*

	Image Subject Ethnic Group		
	EA	**AA**	**CA**
Pain	n.s.	CA & AA*	n.s.
Confidence	CA & AA*	CA & AA*	CA & AA*
Valence	EA & CA*	n.s.	n.s.
Arousal	n.s.	n.s.	n.s.
Dominance	EA & CA** / AA & CA*	n.s.	n.s.

Table 6. EA images: response variables regressed on out-group indicator variables. *(***p<.001; **p<.01)*

Participant Ethnicity				
	CA	**AA**	**F**	**R2**
Pain	n.s.	n.s.	n.s.	
Confidence	-323.99**	-7.423	7.463***	0.1131

Table 7. AA images: response variables regressed on out-group indicator variables. (***p<.01; *<.05*)

Participant Ethnicity				
	CA	**EA**	**F**	**R2**
Pain	-260.83*	-186.46	2.977	0.0484
Confidence	-296.22**	-112.17	4.84	0.0764

Table 8. CA images: response variables regressed on out-group indicator variables. (***p<.01; *<.05*)

Participant Ethnicity				
	AA	**EA**	**F**	**R2**
Pain	n.s.	n.s.	n.s.	
Confidence	249.47**	201.96*	4.34*	0.0691

Given these results, we supported H3a, H3b, and H3c with some interesting caveats. It is clear that the ethnic background of both the participant and the subject depicted in a painful image are correlated to the response variables, and in particular, to self-reported confidence on task. However, rather than providing support for a clear-cut greater in-group sensitivity across all participants, our results highlight differences between our minority participants (EA and AA) and the Caucasian participants. Caucasians report being less accurate in inferring the emotions of both EA and AA sub-

jects. However, unexpectedly, they also report less self-confidence than others, when describing the emotions of other Caucasians in painful settings.

Interestingly, these relationships are mitigated by the degree to which one is in touch with his or her own ethnic identity and background. In particular, we observed that MEIM-ID-search is positively correlated to pain arousal as well as self-reported task confidence. Individuals who have scored high on these items of the MEIM have put forth effort to understand their ethnic background and its impact on their life experiences. This characteristic explains more variance in responses to painful images as compared to in-/out-group relation to the image subject.

The trends concerning the affective content of the emotion tags are less clear. However, it does seem to be the case that ethnic background is relevant, as we observe participants describing EA images using word-tags with differing levels of valence, arousal and dominance, as compared to the EA in-group participants (Table 9).

Table 9. EA Pictures: response variables regressed on out-group member dummies and MEIM scores. (****p<.001; **p<.01; *<.05*)

	AA Participant	CA Participant	ID / Search	Affinity / commitment	F	R2
Pain	47.00	-18.84	33.03*	10.56	4.021**	0.1227
Confidence	-41.639	-190.788	38.904**	-2.561	6.801***	0.1913
Valence	0.01422	0.02860*	0.001617	-0.002425*	2.559*	0.00289
Arousal	0.01957*	0.01819*	0.003492**	-0.002567**	3.696**	0.00417
Dominance	0.01811**	0.01864*	-0.001260	-0.000502	4.613**	0.00520

Table 10. AA Pictures: response variables regressed on out-group member dummies and MEIM scores. (****p<.001; **p<.01*)

	EA Participant	CA Participant	ID / Search	Affinity / commitment	F	R2
Pain	-127.871	-47.626	41.624**	9.266	5.707***	0.1656
Confidence	-60.626	-99.757	39.645**	6.497	6.895***	0.1934

Table 11. CA Pictures: response variables regressed on out-group member dummies and MEIM scores. (****p<.001; **p<.01*)

	EA Participant	AA Participant	ID / Search	Affinity / commitment	F	R2
Pain	-199.415	-66.314	47.088**	1.656	4.179**	0.1269
Confidence	23.184	33.544	53.857***	-9.124	8.106***	0.2199

5 Discussion

As mentioned in the introduction, many believe that increasing the diversity of those involved in all of the processes and tasks that go into building new social technologies – such as automated image tagging – will help ensure they are beneficial for all users. In the current study, crowdsourcing allowed us to gather image metadata on a visual pain emotion recognition task from a diverse workforce, consisting of men and women of several ethnic backgrounds. Our findings support the claim that diversity can be of benefit, but also underscore the need to have access to verified information concerning the personality, such as empathy levels and identities of crowdworkers. This is particularly important for tasks that hinge on one's ability to perceive and interpret the negative feelings of others.

5.1 Interpreting Others' Pain

Two of our response variables quantified our participants' experience on task. The pain arousal rating gauged the extent to which workers were able to feel a depicted subject's pain, while self-reported confidence measured their self-assurance in their ability to describe, using word tags, the depicted subjects' emotion(s).

We found little evidence that worker demographics alone could be used to predict the extent to which one will feel pain for image subjects, or their perceived confidence on task. The one exception here is the correlation between self-reported confidence and ethnicity; Caucasians reported themselves as having less confidence than other ethnic groups, regardless of the ethnicity of the subject depicted.

As compared to demographic characteristics (age, gender, and ethnicity), trait empathy and strength of ethnic identity are more indicative of a participant's ability to perform the task. In particular, those who have high levels of other-oriented empathy (i.e., emotional concern), and who are in touch with their own ethnic background and identity, are likely to be reliable performers on this task. These two variables appear to serve as indicators of one's ability to understand and describe another's feelings of pain and distress.

It is of great importance to better understand the nature of crowdworkers, since the characteristics of MTurkers may be unique and different from the general population. Our findings reveal some differences in the visual pain emotion recognition process of the workers from the general population. The general bibliography indicates significant gender differences in emotion recognition [7; 11; 18; 22; 23; 26; 35; 42; 58; 61], however, our workers do not seem to extend gender differences in the current study of reactivity. This warrants further explanation of how gender identity is affected while performing crowdsourced tasks online, and in some cases, for several hours per day.

5.2 Describing Pain Through Emotion Labels

The remaining three response variables, arousal, valence and dominance of word-tags, quantified three characteristics concerning the affective content of the words participants used to describe the emotions of the individuals depicted in painful images.

Interestingly, while demographic variables proved not to be correlated to participants' level of pain arousal or perceived task confidence, as we expected given the bibliography on gender differences and emotion, they do tell us something about the types of word-tags they might use to describe the emotions of others. For example, women are more likely than men to use labels with higher arousal scores (e.g., describing a woman pictured carrying a child through a flooded area as "panicked" rather than simply "scared"). On the other hand, women are less likely than men to use word-tags suggesting dominance or control (e.g., describing the woman as "defeated" rather than "courageous".) There is a vast literature on gender differences and language, with many suggesting that "women's language" demonstrates their tendency to be more emotional than men, and of course, less powerful, e.g., [30].

There was also evidence suggesting that ethnic background plays a role in the word-tags chosen to describe painful emotions. Interestingly, differences occurred with respect to the tags chosen by EA participants in general (Table 4), as well as words chosen by AA and CA participants to describe images of EA subjects in pain (Table 9). In summary, EA participants use word-tags expressing less dominance or control, in comparison to others. One possible explanation for this is the difference in the emphasis placed on self-expression by various cultural groups [29], which might lead one to use more neutral/forceful language. What is clear here is that recruiting a more diverse workforce for the generation of image metadata, should in turn result in a richer set of image descriptions.

6 Summary and Implications

Our results demonstrate that crowdworkers are not a homogenous group of people, even if they are recruited from within the same country, as in the case of our current study. Their diverse characteristics and the quality of tasks performed should be taken into account when assigning crowdworkers to specific tasks. For instance, we found that gender and age of crowdworkers are correlated to the affective content of words used in the tagging task, but not to workers' pain arousal during the task. To increase the quality of crowdsourcing work, we believe that the nature of the task should be understood clearly, and suggest that a matching algorithm could be used to match tasks with the most relevant workers based on their profiles.

In addition, our findings concerning the correlations between worker demographics (in particular, ethnicity), and the affective content of words they chose to use in their descriptions, have implications for other types of tasks that are commonly crowdsourced. For instance, there is growing interest in using crowdsourcing platforms like MTurk to build resources for natural language processing, including word-level emotion association lexicons [46]. Given the known correlations between demographic characteristics and language use, researchers should carefully consider the nature of the human computation tasks they assign to workers, as well as the characteristics of their workforce. As Law and von Ahn note [32], for many tasks, such as emotion detection and/or association, it may be more reasonable to aim for capturing "cultural truth" rather than "ground truth" (p. 26), in resulting data sets. However, the

question in the case of crowdsourcing is, who's cultural truth are we capturing in the data?

It would therefore be helpful if crowd platforms consider including verified demographic information of workers without compromising their anonymity. Currently, there is limited information of workers' background. Apart from information such as how many tasks they have done, and to what degree of accuracy (as determined by the task "requester"), we know very little about the workers' background. It would be useful if the crowd platforms provide the researcher with basic demographic characteristics, such as gender, age and ethnicity. Furthermore, we found that for some tasks, demographic characteristics do not have any significant correlation to perceived performance. In some cases, workers' personalities matter more than demographic profiles in producing good quality crowd content. Given these findings, a main challenge lies in providing enough worker profile information while still maintaining the individual's anonymity.

7 Conclusion

Our paper sheds light on how crowdworkers interpret emotions through computers and questions the level of empathy the crowd can feel behind the screen. We also examined how crowd diversity is linked to task outcomes. From the results, it is clear that not all crowdworkers are the same, and for certain tasks, we should consider the demographic and personality profiles behind the massive crowd task force to avoid embarrassing and harmful consequences, such as the miss-tagging incident we highlighted at the introduction of this paper. Inclusive design in UI/UX has become an established research/practice area in HCI. We believe that this notion of inclusivity should be extended to crowdsourcing in order to design systems that genuinely "do good".

Therefore, it would be interesting to expand the study in other countries, so we can examine if contextual characteristics beyond ethnicity might affect the empathic process. One of the limitations of Mechanical Turk is that it provides us primarily with workers that are currently living in the United States. Also, it would be useful to study other types of personality characteristics. In our study we focused in the characteristic of empathy through the visual pain emotion judgment process. It would be interesting to expand the current study to examine other personality traits and personality types (e.g., narcissistic personality trait, psychopathy and the Big Five). It is very likely that they will have an impact on the way crowdworkers assess painful emotions.

Furthermore, in this study, we focused only on painful images, depicted humans in distress and as a result a lot of non-verbal information was not available to facilitate the pain emotion judgment process. The inclusion of verbal information may provide the individual with more confidence for emotion judgment. Therefore, in future studies, it would be interesting to study painful emotion judgment through video. Of course, in order to use crowdworkers to assess painful emotion through visual content, we need to consider the privacy of sending images or videos of individuals to the crowd. Future work can focus on how we can obscure one's identity while retaining

key facial and non-verbal characteristics, which can still allow the crowdworkers to accurately classify negative and painful emotions.

References

1. Albas, D. C., McCluskey, K. W., & Albas, C. A. (1976). Perception of the Emotional Content of Speech A Comparison of Two Canadian Groups. *Journal of Cross-Cultural Psychology*, *7*(4), 481-490.
2. Aron, A., Aron, E. N., & Norman, C. (2004). Self-expansion model of motivation and cognition in close relationships and beyond. In M. Brewer, & M. Hewstone, *SElf and social identity* (pp. 99-123). Malden, MA, USA: Blackwell Publishing.
3. Aziz-Zadeh, L., Sheng, T., Liew, S. L., & Damasio, H. (2011). Understanding otherness: the neural bases of action comprehension and pain empathy in a congenital amputee. *Cerebral Cortex, 22(4),* 811-819.
4. Batson, C. D., Turk, C. L., Shaw, L. L., & Klein, T. R. (1995). Information function of empathic emotion: Learning that we value the other's welfare. *Journal of personality and social psychology*, *68*(2), 300 -313.
5. Bolger, E. (1999). Grounded theory analysis of emotional pain. *Psychotherapy Research*, *9*(3), 342-362.
6. Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate?. *Journal of social issues*, *55*(3), 429-444.
7. Brody, L. (2009). *Gender, emotion, and the family*. Harvard University Press.
8. Brown, L. M., Bradley, M. M., & Lang, P. J. (2006). Affective reactions to pictures of ingroup and outgroup members. *Biological psychology*, *71*(3), 303-311.
9. Callison-Burch, C. (2009, August). Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1* (pp. 286-295). Association for Computational Linguistics.
10. Chao, L., Tao, J., Yang, M., Li, Y., & Wen, Z. (2014, November). Multi-scale temporal modeling for dimensional emotion recognition in video. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* (pp. 11-18). ACM.
11. Chaplin, T. M., Cole, P. M., & Zahn-Waxler, C. (2005). Parental socialization of emotion expression: gender differences and relations to child adjustment. *Emotion*, *5*(1), 80-88.
12. Chow, S. M., Ram, N., Boker, S. M., Fujita, F., & Clore, G. (2005). Emotion as a thermostat: representing emotion regulation using a damped oscillator model. *Emotion*, *5*(2), 208-225.
13. Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, *44*(1), 113-126.

14. Dhall, A., Goecke, R., Joshi, J., Wagner, M., & Gedeon, T. (2013, December). Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction* (pp. 509-516). ACM.
15. Ekman, P. E., & Davidson, R. J. (1994). *The nature of emotion: Fundamental questions*. Oxford University Press.
16. Elfenbein, H. A., & Ambady, N. (2002a). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*,*128*(2), 203-235.
17. Elfenbein, H. A., & Ambady, N. (2002b). Is there an in-group advantage in emotion recognition?. *Psychological bulletin*,*128*(2), 243-249.
18. Fischer, A. H., Rodriguez Mosquera, P. M., Van Vianen, A. E., & Manstead, A. S. (2004). Gender and culture differences in emotion. *Emotion*, *4*(1), 87.
19. Golby, A. J., Gabrieli, J. D., Chiao, J. Y., & Eberhardt, J. L. (2001). Differential responses in the fusiform region to same-race and other-race faces. *Nature neuroscience*, *4*(8), 845-850.
20. Goleman, D. (1996). *Emotional intelligence: Why it can matter more than IQ*. London, UK: Bloomsbury.
21. Greenberg, L. S., & Bolger, E. (2001). An emotion-focused approach to the overregulation of emotion and emotional pain. *Journal of Clinical Psychology*,*57*(2), 197-211.
22. Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological bulletin*, *85*(4), 845-857.
23. Hall, J. A. (1990). *Nonverbal sex differences: Accuracy of communication and expressive style*. Johns Hopkins University Press.
24. Hamilton, W. D. (1964). The genetical evolution of social behaviour. *Journal of theoretical biology*, *7*(1), 17-52.
25. Hancock, J. T., Landrigan, C., & Silver, C. (2007, April). Expressing emotion in text-based communication. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 929-932). ACM.
26. Hess, U., Senécal, S., Kirouac, G., Herrera, P., Philippot, P., & Kleck, R. E. (2000). Emotional expressivity in men and women: Stereotypes and self-perceptions. *Cognition & Emotion*, *14*(5), 609-642.
27. Hipp, J. A., Adlakha, D., Gernes, R., Kargol, A., & Pless, R. (2013, November). Do you see what I see: crowdsource annotation of captured scenes. In *Proceedings of the 4th International SenseCam & Pervasive Imaging Conference* (pp. 24-25). ACM.
28. Jackson, P. L., Rainville, P., & Decety, J. (2006). To what extent do we share the pain of others? Insight from the neural bases of pain empathy. *Pain*, *125*(1-2), 5-9.
29. Kim, H. S., & Sherman, D. K. (2007). "Express yourself": culture and the effect of self-expression on choice. *Journal of personality and social psychology*, *92*(1), 1-11.

30. Lakoff, R. (1973). Language and woman's place. *Language in society*, *2*(01), 45-79.
31. Laput, G., Lasecki, W. S., Wiese, J., Xiao, R., Bigham, J. P., & Harrison, C. (2015, April). Zensors: Adaptive, Rapidly Deployable, Human-Intelligent Sensor Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1935-1944). ACM.
32. Law, E., & Ahn, L. V. (2011). Human computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *5*(3), 1-121.
33. Lee, C. M., Narayanan, S., & Pieraccini, R. (2001). Recognition of negative emotions from the speech signal. In *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on* (pp. 240-243). IEEE.
34. Lee, E., Kim, G. W., Kim, B. S., & Kang, M. (2014, November). A Design Platform for Emotion-Aware User Interfaces. In *Proceedings of the 2014 workshop on Emotion Representation and Modelling in Human-Computer-Interaction-Systems* (pp. 19-24). ACM.
35. Lennon, R., & Eisenberg, N. (1987). Gender and age differences in empathy and sympathy. *Empathy and its development*, 195-217.
36. Levenson, R. W., & Ruef, A. M. (1992). Empathy: a physiological substrate. *Journal of personality and social psychology*, *63*(2), 234-246.
37. Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In C.C. Aggarwal & C.X. Zhai (eds.), *Mining text data* (pp. 415-463). Springer US.
38. Luther, K., Pavel, A., Wu, W., Tolentino, J. L., Agrawala, M., Hartmann, B., & Dow, S. P. (2014, February). CrowdCrit: crowdsourcing and aggregating visual design critique. In *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 21-24). ACM.
39. Marshall, V. W. (1996). The state of theory in aging and the social sciences.*Handbook of aging and the social sciences*, *4*, 12-30.
40. Martin, G. B., & Clark, R. D. (1982). Distress crying in neonates: Species and peer specificity. *Developmental psychology*, *18*(1), 3-9.
41. Mathur, V. A., Harada, T., Lipke, T., & Chiao, J. Y. (2010). Neural basis of extraordinary empathy and altruistic motivation. *Neuroimage*, *51*(4), 1468-1475.
42. McClure, E. B. (2000). A meta-analytic review of sex differences in facial expression processing and their development in infants, children, and adolescents. *Psychological bulletin*, *126*(3), 424-453.
43. McDuff, D., El Kaliouby, R., & Picard, R. (2011, November). Crowdsourced data collection of facial responses. In *Proceedings of the 13th international conference on multimodal interfaces* (pp. 11-18). ACM.
44. McVea, C., & Gow, K. (2006). Healing a mother's emotional pain: protagonist and director recall of a Therapeutic Spiral Model (TSM) session. *Journal of Group Psychotherapy Psychodrama & Sociometry -New Series-*, *59*(1), 3-22.

45. Mohammad, S.M. (2015). Sentiment analysis: Detecting valence, emotions, and other affectual states from texts. *Emotion Measurement, 201-238.*
46. Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, *29*(3), 436-465.
47. Morris, R. R., & Picard, R. (2012). Crowdsourcing collective emotional intelligence. *arXiv preprint arXiv:1204.3481*.
48. Morrison, K. (2015, June 09). How Many Photos Are Uploaded to Snapchat Every Second? Retrieved November 14, 2016, from http://www.adweek.com/socialtimes/how-many-photos-are-uploaded-to-snapchat-every-second/621488
49. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1-135.
50. Phinney, J. S. (1992). The multigroup ethnic identity measure a new scale for use with diverse groups. *Journal of adolescent research*, *7*(2), 156-176.
51. Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *23*(10), 1175-1191.
52. Preston, S. D., & De Waal, F. B. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and brain sciences*, *25*(01), 1-20.
53. Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (2010, June). Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 139-147). Association for Computational Linguistics.
54. Robb, D. A., Padilla, S., Kalkreuter, B., & Chantler, M. J. (2015, April). Crowdsourced Feedback With Imagery Rather Than Text: Would Designers Use It?. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1355-1364). ACM.
55. Ruffman, T., Henry, J. D., Livingstone, V., & Phillips, L. H. (2008). A meta-analytic review of emotion recognition and aging: Implications for neuropsychological models of aging. *Neuroscience & Biobehavioral Reviews*,*32*(4), 863-881.
56. Ryff, C. D., Magee, W. J., Kling, K. C., & Wing, E. H. (1999). Forging macro-micro linkages in the study of psychological well-being. *The self and society in aging processes*, 247-278.
57. Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social science information*, *44*(4), 695-729.
58. Scherer, K. R., Wallbott, H. G., & Summerfield, A. B. (1986). *Experiencing emotion: A cross-cultural study*. Cambridge University Press.
59. Schroeder, D. A., Dovidio, J. F., Sibicky, M. E., Matthews, L. L., & Allen, J. L. (1988). Empathic concern and helping behavior: Egoism or altruism? *Journal of Experimental Social Psychology*, *24*(4), 333-353.
60. Schwartzberg, S. S., & Janoff-Bulman, R. (1991). Grief and the search for meaning: Exploring the assumptive worlds of bereaved college students. *Journal of Social and Clinical Psychology*, *10*(3), 270-288.

61. Simon, R. W., & Nath, L. E. (2004). Gender and Emotion in the United States: Do Men and Women Differ in Self-Reports of Feelings and Expressive Behavior? 1. *American journal of sociology*, *109*(5), 1137-1176.
62. Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008, October). 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* (pp. 254-263). Association for Computational Linguistics.
63. Tao, J., & Tan, T. (2005). Affective computing: A review. In *Affective computing and intelligent interaction* (pp. 981-995). Springer Berlin Heidelberg.
64. Triandis, H. C., Bontempo, R., Villareal, M. J., Asai, M., & Lucca, N. (1988). Individualism and collectivism: Cross-cultural perspectives on self-ingroup relationships. *Journal of personality and Social Psychology*, *54*(2), 323-338.
65. Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly review of biology*, 35-57.
66. Vinacke, W. E., & Fong, R. W. (1955). The judgment of facial expressions by three national-racial groups in Hawaii: II. Oriental faces. *The Journal of Social Psychology*, *41*(2), 185-195.
67. Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, *45*(4), 1191-1207.
68. Wolfgang, A., & Cohen, M. (1988). Sensitivity of Canadians, Latin Americans, Ethiopians, and Israelis to interracial facial expressions of emotions. *International Journal of Intercultural Relations*, *12*(2), 139-151.
69. Xu, A., Huang, S. W., & Bailey, B. (2014, February). Voyant: generating structured feedback on visual designs using a crowd of non-experts. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 1433-1444). ACM.
70. Zahn-Waxler, C., & Radke-Yarrow, M. (1990). The origins of empathic concern. *Motivation and emotion*, *14*(2), 107-130.