It's About Time: A View of Crowdsourced Data Before and During the Pandemic

Evgenia Christoforou
Research Centre on Interactive Media, Smart Systems and Emerging Technologies
Nicosia, Cyprus
e.christoforou@rise.org.cy

Pınar Barlas
Research Centre on Interactive Media, Smart Systems and Emerging Technologies
Nicosia, Cyprus
p.barlas@rise.org.cy

Jahna Otterbacher
Open University of Cyprus &
Research Centre on Interactive Media, Smart Systems and Emerging Technologies
Nicosia, Cyprus
jahna.otterbacher@ouc.ac.cy

ABSTRACT

Data attained through crowdsourcing have an essential role in the development of computer vision algorithms. Crowdsourced data might include *reporting biases*, since crowdworkers usually describe what is "worth saying" in addition to images' content. We explore how the unprecedented events of 2020, including the unrest surrounding racial discrimination, and the COVID-19 pandemic, might be reflected in responses to an open-ended annotation task on people images, originally executed in 2018 and replicated in 2020. Analyzing themes of *Identity* and *Health* conveyed in workers' tags, we find evidence that supports the potential for temporal sensitivity in crowdsourced data. The 2020 data exhibit more race-marking of images depicting non-Whites, as well as an increase in tags describing Weight. We relate our findings to the emerging research on crowdworkers' moods. Furthermore, we discuss the implications of (and suggestions for) designing tasks on proprietary platforms, having demonstrated the possibility for additional, unexpected variation in crowdsourced data due to significant events.

CCS CONCEPTS

• **Information systems** → **Computing platforms**; • **Human-centered computing** → *Empirical studies in HCI*; • **Computing methodologies** → *Artificial intelligence.*

KEYWORDS

crowdsourcing, data reproducibility, image annotation, reporting bias, temporal sensitivity

ACM Reference Format:
Evgenia Christoforou, Pınar Barlas, and Jahna Otterbacher. 2021. It's About Time: A View of Crowdsourced Data Before and During the Pandemic. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3411764.3445317

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445317

1 INTRODUCTION

Paid micro-task crowdsourcing is an essential tool in the creation of datasets for training and evaluating computer vision algorithms. In particular, datasets designed for object and scene recognition typically require *descriptive labeling* at some stage in the pipeline. For example, the ImageNet project [15] is reported to be the largest academic user of Amazon's Mechanical Turk, employing between 20-30k workers per year to label and/or verify semantic image labels.[1] Similarly, Microsoft COCO is another widely-used dataset, which relies on MTurkers for labeling and instance detection (i.e., detecting objects and scenes) [23], as well as for providing descriptive captions for images [7].

While the availability of such large, human-enriched datasets has been a boon to computer vision research, there is increasing awareness of the human biases that are reflected in crowdsourced data. Dumitrache rejected the notion that there can be a single ground truth in any semantic annotation task, arguing instead for a "disagreement-aware" approach to crowdsourcing [17]. In a similar vein, Chung and colleagues [8] noted the diverse answers often provided by workers, and advocated for reporting statistical distributions of responses, to preserve this diversity. In explaining the variation in responses to open-ended image annotation tasks, Berg and colleagues [3] described crowdsourced annotations as being fundamentally "human-centric"; when we ask workers to describe images, they provide us with a wealth of social information beyond that for which there is direct visual evidence. Misra and colleagues [26] further developed this view, emphasizing that workers' annotations on images should be not considered as a faithful description of content, but rather, constitute a report of what is "worth saying" about the target image.

A concrete example of the human biases found in image annotations is that of gender and racial stereotypes. In his analysis of the Flickr30K dataset, van Miltenburg [44] found that workers make inferences on images of people, which do not logically follow from the image content. In particular, he noted cases of gender, racial and ethnic stereotypes, and other "unwanted inferences" (e.g., ethnicity marking, suggesting that images of White people are the default). Zhao and colleagues noted rampant gender biases in the MS COCO dataset [47], as a result of the types of images included in the dataset, but also the annotations accompanying them, leading

[1] https://www.nytimes.com/2012/11/20/science/for-web-images-creating-new-technology-to-seek-and-find.html

to bias amplification in the trained computer vision models. Hendricks and colleagues reported similar findings in the MS COCO data for an image captioning task [21]. Finally, gender stereotyping has also been documented in data arising through a gamified image labeling task, the ESP Game [30], in which players were more likely to describe physical traits for images of women (as compared to men), and the inferred professional roles for images of men (as compared to women). In summary, it is evident that crowdsourcing, and in particular image annotation tasks, are key elements of the process by which society's biases are amplified and perpetuated in computer vision algorithms.

While the above constitute examples of datasets built and used mainly by experts, it is becoming increasingly easy and economical for non-experts to create datasets for computer vision and other machine learning tasks, that better serve their needs. The recent industry trend of "democratizating artificial intelligence," together with the rise in "no-code tools" and "cognitive services" are empowering non-expert business users.[2] Data harvesting platforms – powered by crowdwork – have also followed suit. While MTurk has always provided easy-to-use templates for a range of human intelligence tasks (HITs), crowdsourcing platforms such as Appen[3] and Clickworker[4] took an additional step by offering a full range of data collection and "crowd management" services. Notably, computer-vision start-up Clarifai[5] now offers integrated end-to-end services, from dataset creation and labelling to training and evaluating models. Despite the convenience of such services and their potential to aid innovation, there is awareness that they may be a "double-edged sword," given the prevalence of biases which may go unrecognized by non-experts.[6]

Given the above concerns, it is becoming increasingly important to better understand the nature of human biases that can manifest in data crowdsourced from popular platforms and services, and the extent to which particular characteristics of the platform and/or task may exacerbate them. As will be explained, in the current work, we study the *temporal effects* that significant global and local events might have had on an open-ended task involving the description of people-related images.

2 MOTIVATION

To examine temporal effects, we replicate an open-ended image annotation task, which was originally designed and carried out two years prior, comparing the data received in 2020 versus that collected in 2018. Next, we will motivate the choice of our crowdsourcing task. Following that, we explain why we expect to observe temporal sensitivity, as a result of the influence of 2020's significant events on crowdworkers' responses.

As previously discussed, human-labeled training datasets for computer vision algorithms often suffer from social bias, which is most evident in the annotations on people-related images. This bias is, of course, carried down the development pipeline and is typically reflected in the final output of the algorithms trained on the data. There is a growing body of research documenting the consequences of this in the social world (e.g., the algorithmic misgendering of images of women and people of color [22] and of non-binary people [37]). Thus, bias in computer vision training datasets can have great implications. While others have noted the issue of imbalanced representation in training data (e.g., [6, 47]), we focus on the manner in which images in training data are annotated by crowdworkers. In particular, from the works discussed in the Introduction, it appears that popular, open-ended crowdsourcing task designs give a higher degree of freedom to the annotators to express what they find "worth saying" in a image and thus, potentially introduce their own biases into the dataset. This may be especially true when the requester of the task is a non-expert, and simply uses a platform's template uncritically. Thus, an *open-ended image annotation task* is a good candidate for our study on temporal sensitivity, as it leaves room for workers' interpretations on the image. At the same time, it represents a common task in building high-impact computer vision datasets.

Having motivated our crowdsourcing task, we now discuss the importance of studying temporal sensitivity, in light of the current world climate. The unprecedented events of 2020 have created a temporary social imbalance that has influenced people around the globe. Thus, we take advantage of this "organic" and universal change in the crowdworkers' environment and study the temporal variation manifested in the collected data. We hypothesize that among other factors, the difficulties experienced worldwide, given the COVID-19 pandemic, coupled with the continuing social unrest surrounding racial discrimination in the U.S., have influenced the datasets generated through crowdsourcing. In particular, given the significant impact of these events and their extensive media coverage, we expect that concepts central to these events (i.e., issues of *health* and *identity*) are "at the forefront of thought" [29]. This hypothesis is in parallel to findings from cognitive psychology, which demonstrated that the public's perceptions of *airplanes* changed after the September 11th terrorist attacks. While a causal effect could not be definitively claimed based on the experimental results reported by Novick [29], she postulated that the change in people's perceptions of airplanes – as being more "typical" examples of the semantic category "vehicle" after 9/11 as compared to before – was likely due to the frequent exposure to media coverage of the attacks. Similarly, the current work studies how time influences crowdsourced data, by showing a connection between the present events and the data collected from a crowdsourcing task in 2020, as compared to the data collected from the same task in 2018.

Public health researchers have described how these unprecedented experiences have affected our well-being. Individuals may be suffering from confusion, isolation, and feelings of insecurity, while communities are facing secondary crises brought about by the lack of resources for medical response, schooling and childcare, among other challenges. [34]. Large-scale problems are being reported, such as alcohol and drug abuse [9] as well as increased levels of anxiety and sleep disturbances in the general population [41], together with some early signs of eating disorders [19] and the danger of weight gains over short periods of time [5]. For some individuals with existing eating disorders, the pandemic has been

[2]https://www.forbes.com/sites/forbesfinancecouncil/2020/08/11/a-no-code-environment-brings-ai-to-the-business-user/
[3]https://appen.com/
[4]https://www.clickworker.com/
[5]https://www.clarifai.com/
[6]https://www.strategy-business.com/article/Democratizing-artificial-intelligence-is-a-double-edged-sword?gko=ffdcd

a trigger for increased anxiety and worsening symptoms [43]. In the midst of all this, the "stay at home" movement and/or enforced lockdowns, in combination with the economic crisis, have created a fruitful environment for crowdwork supply to bloom and demand to increase,[7] with on-premise laboratory studies being suspended in many areas [36]. A plethora of new datasets produced through crowdsourcing are being created, but can crowdwork during a pandemic yield reliable data?

Thus, we seize the opportunity to observe crowdworkers during what clearly cannot be considered "normal times," providing evidence that *societal events* can introduce temporal variations in resulting data. We leverage an existing image annotation dataset [2], created in December 2018 through a generic task presenting workers with standardized, passport-style images of people from the Chicago Face Database [25]. We replicate the image annotation task as described in [2] using the same platform. As will be detailed, we find that themes of *Identity* (words relating to race and nationality) as well as *Health* (e.g., body weight) are used significantly more frequently by workers to describe images in 2020, as compared to the 2018 data, supporting our hypothesis that the events of 2020 may have contributed to the variations in the resulting data. We discuss the challenges of crowdsourcing during a time of heightened stress, as well as the need to cope with temporal effects. Finally, we provide a set of guidelines for accomplishing this.

3 BACKGROUND AND RELATED WORK

In this section, we take a closer look at the value of the information to be gained by replicating an open-ended image annotation task. In addition, we ground the approach used in analyzing the workers' descriptions of the images.

3.1 Stress, mood and (crowd)work

Although it is clear that the present circumstances have brought about significant stress and uncertainty to much of the population, we are only beginning to understand how this has affected work, both in terms of process and outcomes. Early studies have considered the particular stresses of healthcare workers [39] or of parents returning to work with schools closed [11]. Likewise, researchers are considering the complex interactions between the Black Lives Matter protests, their psychological impact on racial minorities [45] as well as public efforts to enforce social distancing to combat the spread of COVID-19 [14]. Park and colleagues [33] surveyed MTurkers to understand their stress during this time period, as well as their coping mechanisms. However, we are unaware of published studies that have considered how the current situation has impacted the process and outcomes of *crowdwork* in particular.

As noted by Zhuang and Gadiraju [48], while there is a substantial body of literature that links workers' moods to processes and outcomes in the physical workplace, little is known about mood and crowdwork. Therefore, they considered the relationship between crowdworkers' moods and their perception of work as well as their performance on task, conducting a two-phase study. First, a survey suggested that workers' moods correlated to their perceptions of their level of engagement with a task; however, a follow-up experiment involving an information-finding HIT demonstrated no significant correlation between performance (outcome) and self-reported moods. In contrast, other researchers have attempted to harness moods to enhance crowdwork. For instance, Morris and colleagues [27] found that inducing a happy mood through playing music, helped workers perform better on creative tasks. Finally, Shen et al. [40] suggested an "affective crowdsourcing" approach, which attempted to account for workers' moods when scheduling tasks, in order to maximize their collective efficiency.

These findings, along with reports of generalized feelings of stress and anxiety during the events of 2020 – and even boredom and frustration during lockdown [16] when the population of crowdworkers likely increased [24] – suggest that we may find differences in the data produced in 2020, as compared to that collected during a less stressful time period. Although the literature to date does not evidence a direct link between workers' stress or mood during the pandemic, to the quality of their work, it is logical to anticipate an observable, temporal sensitivity, based on the unique characteristics and significance of this period of time.

3.2 Image annotation as communication: analyzing textual responses

The micro-task we conducted asks workers to "translate" the rich visual, but implicit, information contained in an image, by providing an explicit encoding of that information (i.e., word tags). Thus, the task is fundamentally a communication process; this explains why social information is conveyed in workers' responses, as demonstrated in previous research [3, 26]. In particular, the way we express ourselves in writing reveals a great deal of (often unintended) information, such as our emotional state, our membership within social group(s) as well as our relationships to others [42].

For instance, stylistic changes in writing can often be observed after the experience of emotional trauma; thus, textual analysis has been used to explore collective and individual traumatic experiences. Cohn and colleagues [10] conducted a large-scale analysis of U.S. users of an online diary service, comparing their writing before and after the events of September 11th. They noted changes in the emotions expressed in text, as well as changes in the use of stylistic markers, such as pronouns. In particular, after the terrorist attacks, writers were more often socially distanced, which decreased over time (as indicated by increasing use of first-person pronouns over time, which show personal engagement). Similar trends were noted in a Spanish study of online expressive writing after the March 11th terrorist attacks [18].

While the above findings pertain to expressive writing, social psychologists have also considered the characteristics of the language used for a specific purpose – describing other people. First, it is clear that people do not only describe how they literally see another person; they also make inferences about others' abstract characteristics and traits. Even without contact (e.g., when shown a photo), this happens automatically and almost immediately [46]. Fiske and Cox compared the concepts used to describe a friend versus a stranger, finding that strangers were more often described with physical attributes, while familiar persons were more likely described in an interpretive manner [20]. Semin and colleagues

[7] https://cedefop.europa.eu/en/news-and-press/news/has-coronavirus-crisis-made-us-all-crowdworkers

detailed a distinction between the use of abstract/inferential versus concrete language [38]. While the former is not based on visual evidence, the latter follows logically and directly from visual evidence. Furthermore, abstract language (e.g., describing someone as "intelligent" or "kind") implies stability over time and generalizability across situations, whereas concrete descriptions (e.g., noting the clothing worn or the facial expression) report an observation "here and now." Psychologists have noted a strong tendency for us to describe more expected or stereotype-congruent people more abstractly, as compared to less familiar / unexpected people. This tendency is known as *linguistic bias* and is believed to play a role in the creation and perpetuation of social stereotypes [4].

3.3 Research questions

Inspired by the above findings, we analyze crowdworkers' descriptions of the people images, considering the use of two themes in their chosen word tags: concepts related to *Health* and racial or national *Identity*. Given the tendency for traumatic events as well as social relations to influence the use of language, we expect to observe differences in the use of these themes across time. Finally, we also consider the use of inferential (abstract) versus concrete tags to describe people, given their correlation to the social relationship between the perceiver (i.e., worker) and the person being described, in light of the ongoing racial tensions. In particular, we answer the following research questions:

RQ1. Do workers in 2020 refer to the identity and health of the depicted persons, more so than in 2018?

RQ2. Do workers use abstract/inferential versus concrete tags with similar frequencies over time?

4 METHOD

4.1 Data collection and pre-processing

We replicated an image annotation task, originally executed by Barlas et al. [2] on the Appen[8] platform. Their task was designed to emulate the common open-ended image tagging templates at platforms such as MTurk and Appen. The researchers used a highly standardized set of people images, the Chicago Face Database (CFD) [25] featuring individuals from four racial groups, which were self-reported. As shown in Figure 1, the individuals depicted wear the same grey t-shirt, have a neutral facial expression, and directly face the camera. The CFD is composed of 597 images in total: 109 depict Asian, 197 depict Black, 108 depict Latino/a and 183 depict White persons, balanced by gender within each group. Contrary to [2], which targeted India and U.S.-based workers in separate runs, we are only interested in participants based in the U.S. Thus, we ran an identical task with the same images, restricted to crowdworkers registered in the U.S., similarly asking for three unique judgements per image. Crowdworkers were presented with an image, and asked simply to describe its "content" through 10 tags (consisting of one to two words) of their choice. Workers were permitted to describe up to 20 images. They were compensated 0.30 USD per image, with the mean time on task being 120 seconds. We adopted all the same quality control measures described in [2]. In 2020, our study was active from May to June, while the study presented in [2] was active during December 2018 (i.e., 18 months apart).

For the purpose of comparing the data produced in the two studies, we used a portion of the 2018 dataset [2] which is freely accessible. [9] In both the 2018 and 2020 datasets, there are a small number of responses in Spanish. Contrary to the 2018 study, we translated the Spanish-language tags – both those found in the 2018 dataset as well as those collected in our 2020 study – into English. We used the same spell-check process as described in the 2018 study (on both the tags collected in 2020, and all the newly-translated Spanish-language tags) in order to fully replicate the process.

4.2 Thematic tag clusters

In the 2018 study, all processed tags were grouped into clusters following a specific typology (see [2], Table 4). However, for the needs of this work, we isolated the tags describing a person's *Health* and *Identity* appearing in the tags from both years (see Table 1 for examples), creating new clusters for our analysis. As detailed in Figure 2, we consider tags that describe aspects of a person's *Health* and *Identity*, as expressed by the crowdworkers, which make up the two themes of our current typology of tags. Within the theme related to a person's *Health*, we find two respective clusters {Health, Weight}. Since Weight may be an indicator of health, we include these tags under the *Health-related* theme[10]. The Symptoms sub-cluster contains all the tags that could be used to describe a physical feature that may indicate a health condition, while the Overall Health sub-cluster contains all the tags that describe whether a person is generally healthy or not. Tags referring to a physiological characteristic of a person that might be otherwise related to a health condition (e.g., "albino", "broken-nosed") were not included. Furthermore, tags referring to the color of a specific body part (i.e., "pale_face") were also excluded for being too ambiguous. The Weight cluster consists of three sub-clusters: {Overweight, Underweight, Normal Weight}. Tags describing the person's body structure (e.g., "heavy_build") were excluded.

Similarly, the *Identity* theme includes tags used by the crowdworkers to specify a characteristic of a person's physical appearance that could potentially be used to identify a person as belonging to a certain nationality, ethnicity or race. The *Identity* theme contains three clusters: {External Features, Nationality, Race}. The External Features cluster contains the sub-clusters describing a person's skin tone (see Figure 2) and physical features that are described with inferences to the person's ethnicity, race, or nationality (e.g., "chinese_eyes," "latin_skin"). The Nationality cluster consists of sub-clusters of the nationalities most often reported along with the Other sub-cluster, which refers to nationalities appearing only a few times (e.g., "Peruvian" or "Korean"). Finally, the Race cluster includes four races common in the U.S. (which are analogous to the race categories used in the CFD) as well as the sub-cluster Multi-race, consisting of tags such as "half_black" and "asian_black."

Note that in Figure 2, the sub-clusters have been coded as to whether they represent inferential or concrete characterizations of the target person. For example, tags referring to the Overall Health

[8] https://appen.com/ - The 2018 study used the FigureEight platform, later acquired by Appen.

[9] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/APZKSS

[10] We recognize that weight is a ubiquitous topic. We only place weight relevant tags in the Health category with the intention of placing these tags into context with Identity relevant tags.

Figure 1: Four images from the Chicago Face Database (CFD) (left to right: AM-253, BF-233, LM-220, WF-036).

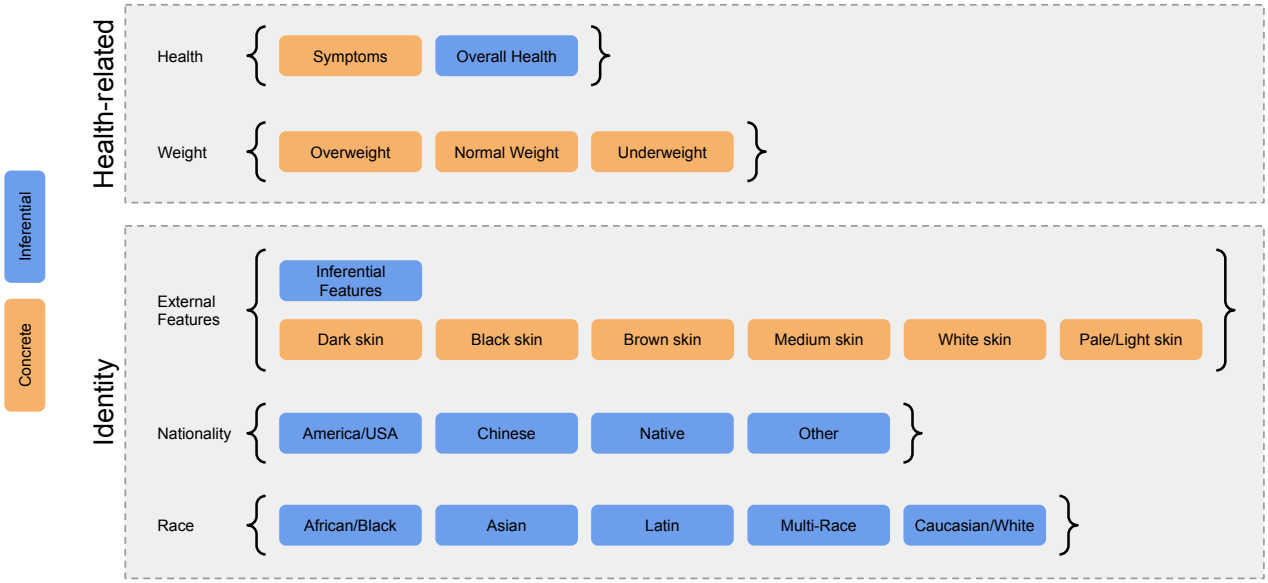

Figure 2: Event-driven themes: *Health-related* and *Identity*, and their respective sub-clusters.Color-coding indicates whether a cluster is inferential or concrete. The sub-clusters, Symptoms, Overweight, Normal Weight and Underweight and the sub-clusters describing a persons skin belong to the concrete characterization; the rest of the sub-clusters belong to the inferential characterization. Clusters are mutually exclusive.

of a person ("healthy," "fit") are considered *abstract/inferential*, as they cannot be ascertained directly based on visual evidence alone. In contrast, the sub-cluster Overweight, consisting of tags such as "heavy" or "plump," is *concrete*, as its tags are based on visual evidence.

5 DATA ANALYSIS

Our analysis focuses on the workers' use of tags that refer to the topics of Health and Identity. Therefore, tags analyzed will either belong exclusively to one of the (sub-)clusters under Health or Identity, or will be from the set of "other remaining tags" that are not processed in this work (see Section 6.2). Each of the 597 images received three judgments, each producing 10 tags. However, as part of the pre-processing, tags that were considered low-quality as described in [2], were excluded. Thus, in our analysis, we consider a total sample size of 17,699 tags in 2018 and 17,224 tags in 2020. Breaking this down by the race of the depicted person, images with a White subject received in total 5,413 tags in 2018, and 5,216 tags in 2020; Black subject images received 5,873 tags in 2018 and 5,720 in 2020; Asian subjects a total of 3,199 tags in 2018 and 3,139 tags in 2020, and Latino/a subjects a total of 3,214 tags in 2018 and 3,149 tags in 2020.

For each tag belonging to either dataset (i.e., 2018 v. 2020), we computed the number of unique occurrences.[11] By the term *unique occurrences* we mean that in the event a worker provided the same tags for the same image more than once, we count it as a single occurrence. Of course, if the same tag appeared in two different judgments, it is counted twice. Thus, the number of unique occurrences in our analysis will inform us of how many tags we have for a particular (sub-)cluster, counting only the repetitions of a tag that do not belong to the same worker, for the same judgement. From this point onward, we refer to the number of unique occurrences simply as *occurrences* of a tag or *# occurrences*. In Section 5.1, we present a thematic perspective on the use of tags, comparing the

[11] We were granted access, by the authors of [2], to the raw data of the U.S. workers in order to count the number of unique occurrences in the 2018 study.

Table 1: Example of tags belonging to the *Health-related* and *Identity* themes, and their respective sub-clusters.

(Sub-)Cluster	Example tags	Sub-cluster	Example tags
Health			
Symptoms	*pain, dehydrated, sweaty*	Overall Health	*healthy, cadaverous, fit*
Weight			
Overweight	*overweight, fat, obese*	Normal Weight	*normal_weight, average_weight*
Underweight	*thin, skinny, skeletal*		
External Features			
Inferential Features	*chinese_eyes, asian_look*	Medium Skin	*medium_skin_color, olive_skin*
Dark Skin	*dark_skin, dark_complexion*	White Skin	*white_skin, face_white*
Black Skin	*black_features, black_skin*	Pale/Light Skin	*fair_complexion, light_skin*
Brown Skin	*brownish_skin, face_brown*		
Nationality			
American/USA	*american, usa, afro-american*	Native	*native_woman, native_american*
Chinese	*china, chinese_man*	Other	*mexican, vietnam, non_american*
Race			
African/Black	*black_guy, afro-descendant, african*	Multi-race	*half_black, multi_race*
Asian	*asian, asia, asiatic*	Caucasian/White	*white_girl, caucasian, white_person*
Latin	*latino, hispanic, latin_girl*		

use of *Health-related* tags across the 597 people images, across time. Moreover, we consider the use of *Identity-related* tags at both points of time, as well as by the race group of the persons depicted in the target images. Following that, in Section 5.2, we present the stylistic perspective on the workers' use of tags, considering where they use more abstract or concrete tags to describe the depicted individuals, and if these stylistic tendencies have changed over time.

5.1 Thematic perspective: *Health* and *Identity*

We begin with a high-level look at the use of tags belonging to the topics of Health and Identity, applying a chi-squared test to examine whether the number of occurrences of the types of tags workers provided (i.e., Health, Identity, or Other remaining tags) are independent of the year these tags were collected, which is our null hypothesis. As expected, the observed distributions of tags by topic differ between 2018 and 2020 ($X^2(2, N = 34923) = 14.57$, $p < .001$). In order to provide answers to our research questions, we take a closer look at the extent to which workers describe the health and identity of the depicted persons, analyzing separately the two topics.

5.1.1 Health-related tags. We now focus on the use of the Health-related tags. As shown in Figure 3, we observe an increase in 2020 in the use of tags belonging to the Overweight, Underweight and Symptoms sub-clusters. We apply a two proportion z-test to see whether the proportion of tags used that are *Health*-related (i.e., # occurrences of Health-related tags over the total sample size of the tags, as presented above) is the same across the two years. Since we are examining four hypothesis at the same time (one for each of the *Health*-related sub-clusters) we apply a Bonferroni correction, adapting our significance level to $0.05/4 = 0.0125$. For the Overweight sub-cluster, the observed difference between 2018 and 2020 is significant ($z = -3.288$, $p = .001$). On the other hand, for the Underweight sub-cluster, the results are not significant ($z = -2.0913$, $p = .036$). Notice that we consider a strict significance level in our analysis; thus, results for the Underweight sub-cluster could be considered significant under different conditions. From this point forward we will report only the exact values for $p < .05$ together with the Bonferroni correction we apply in each case, clearly mentioning the significance level we consider.

In Figure 3, we omitted the graphical representation of the Normal Weight sub-cluster, since we observed only four unique tag occurrences from that sub-cluster in 2018 and none in the 2020 study. A total of 27 images were described in both 2018 and 2020 with a tag belonging to the Overweight sub-cluster. In terms of images tagged at only one point in time, there were 50 distinct images receiving a tag belonging to the Overweight sub-cluster in 2020, as compared to only 16 in 2018. Finally, our analysis indicates that workers provided a similar number of tags describing the Symptoms ($z = -0.42$, $p > .05$) and Overall Health ($z = 0.88$, $p > .05$) of the depicted person in both years.

The above observations address **RQ1**, as far as *Health-related* tags are concerned, with a ***positive*** answer. It is somewhat surprising to see that tags belonging to the Overall Health and Symptoms sub-clusters were not more frequently used in 2020, given the events of the pandemic. Again, parallel to [29], our initial assumption was that with the pandemic and issues of personal and public health being discussed extensively in the media, Health-related tags would be used more frequently. Interestingly enough though, crowdworkers in the 2020 study did not increase their use of Overall Health and Symptoms tags but instead, used significantly more tags belonging to the Weight cluster, in comparison to 2018. By posing **RQ1**, we anticipated, up to a certain degree, that workers' descriptions would also reflect their concerns and stresses related to the pandemic. Our results indicate that weight-related issues are on their minds. According to Google Trends,[12] during April-May 2020, there was a spike in search terms such as "recipes" and "quarantine workout," reinforcing the idea that such topics were on people's minds. In other words, workers may be most likely to notice and mention – when describing an image – what is directly affecting them at the present moment, e.g., weight-related concerns resulting from the circumstances of the pandemic. Of course, the use of more weight

[12] trends.google.com/trends/explore?geo=US&q=quarantine%20workout,home%20recipes

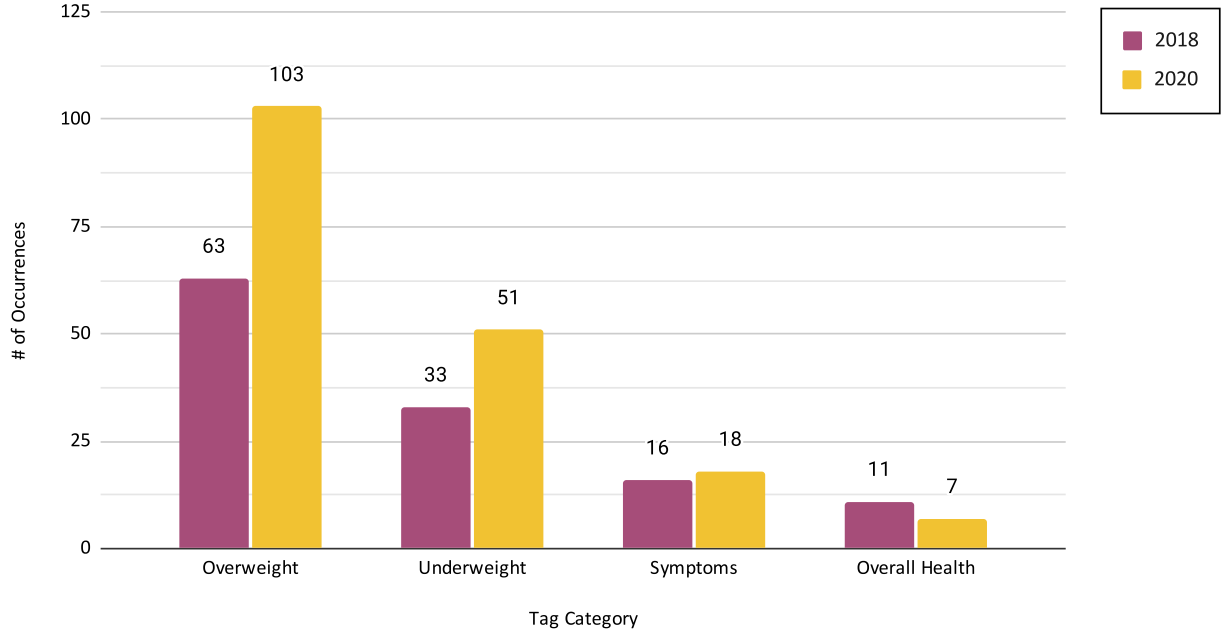

Figure 3: Number of tag occurrences over time, for sub-clusters under Health and Weight clusters. For each sub-cluster, left bar represents the year 2018 and the right bar the year 2020.

related tags in 2020 might also be correlated with other factors affecting the crowdworkers and their perception of the subjects, of which we are unaware.

5.1.2 Identity-related tags. Next, we explore the workers' use of tags related to the depicted person's identity, in addressing **RQ1**. Figure 4 shows the observed tag occurrences of the three clusters of *Identity*-related tags {External Features, Race, Nationality} and their respective sub-clusters. Over all images, workers are approximately three times more likely to mention the person's External Features and Race than to infer the person's Nationality.

Considering the External Features cluster, a two proportion z-test reveals that the proportion of tags used (i.e., the number of occurrences over the total sample size) for the two years we consider is similar ($z = 0.81$, $p > .05$), thus providing a ***partially negative*** answer to the *Identity*-related aspect of **RQ1**. Furthermore, we notice that for all races of the persons depicted in the images, the number of tag occurrences in 2018 versus 2020 are similar. In other words, there are no statistically significant differences across time, in the use of these types of tags when referring to a White subject ($z = 1.27$, $p > .05$), when describing a Black subject ($z = 0.92$, $p > .05$), for tags referring to Asian subjects ($z = 0.38$, $p > .05$) or when describing Latino/a subjects ($z = -1.41$, $p > .05$). Notice that the sample sizes when considering the different proportions are different based on the year and the race of the depicted subject in the image as described in the introduction of this Section.

Looking at the different sub-clusters of the External Features tags, it is interesting to notice that in 2018, there is a strong preference from workers in tagging people as Pale/Light colored, while in 2020, there is a shift towards tagging them as having White Skin/features. In fact, considering a two-proportion z-test for the White sub-cluster tags, we observe a significant difference in occurrence between the two years ($z = -7.35$, $p < .001$). A similar trend is noticeable for describing people of color. While in 2018, significantly more tags in the Dark Skin sub-cluster were observed ($z = 5.12$, $p < .001$), in 2020, we have significantly more tags referring to Black Skin ($z = -4.05$, $p < .001$) or Brown Skin ($z = -2.9982$, $p = .003$). In other words, more specific tags are observed in 2020 to describe a person's color. Notice that we compared together the four hypotheses relevant to the External Features sub-cluster tags, presented above; thus, our threshold for significance is $p < 0.0125$, according to the Bonferroni correction.

It appears that tags used by the workers describing External Features referring to the colors White and Black/Brown are trending in the 2020 dataset, while in 2018, the sub-clusters Pale/Light and Dark were more frequently used. Among the most popular dictionary words used to write about the Black Lives Matter (BLM) movement are the words "white" and "black,"[13] thus serving as evidence that workers might be subject to attentional bias. Of course, it must be noted that this is one among many factors that could affect such behavior.

Looking at the cluster of Race-related tags, we notice no statistically significant increase in the total number of tag occurrences between the two years ($z = -0.92$, $p > .05$); thus again, a ***partially negative*** answer is given regarding the *Identity*-related part of **RQ1**. However, we observe a significantly larger number of Race-related tag occurrences in 2020 describing images depicting Black subjects ($z = -6.04$, $p < .001$). Moreover, we observe the opposite behavior regarding Latino/as ($z = 2.53$, $p = .011$) and

[13] https://home.oxfordowl.co.uk/blog/500-words-black-lives-matter-how-are-british-children-responding-to-the-emerging-themes-and-issues-in-their-writing/

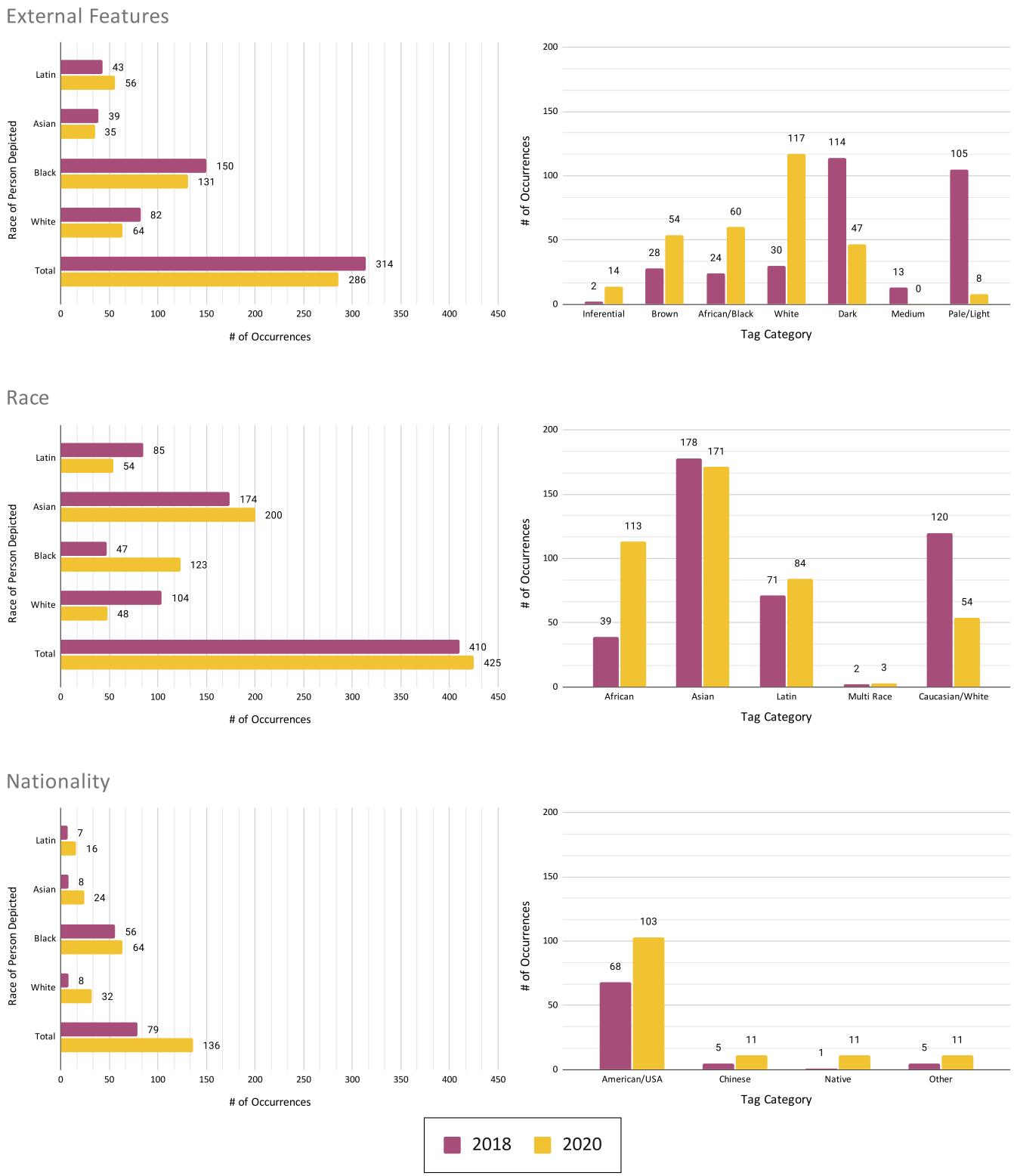

Figure 4: The number of occurrences of tags for the clusters External Features, Race, and Nationality, in 2018 (left/top bar) v. 2020 (right/bottom bar), broken out by the race of the person depicted (left col.) and the specific tag sub-clusters (right col.).

White subjects ($z = 4.34$, $p < .001$), with significantly more Race-related tag occurrences in 2018 as compared to 2020. Notice that the proportion of tags for images depicting Asian subjects are similar among the two years ($z = -1.57$, $p > .05$). Considering the we are comparing four different hypotheses above, we set the threshold for significance to $p < 0.0125$, following the Bonferroni correction.

The above observations relevant to the race of the depicted subject are in line with the observations on the sub-clusters on the Race-related tags. For instance, in 2020, we find workers using a significantly larger number of tags belonging to the African/Black sub-cluster with ($z = -6.18$, $p < .001$), while the tags from the Caucasian/White sub-cluster are significantly more frequent in 2018 with ($z = 4.83$, $p < .001$). Previous research on crowdwork and image labeling [32, 44] reported race-marking (i.e., greater tendency to use race-related words when describing non-White individuals). Those observations are partly in line with the above results. The Asian images in the CFD included individuals with South Asian heritage (i.e., who didn't "look East Asian/Chinese"), hence this could be one factor contributing to our not finding anti-Chinese sentiments in the workers' responses, despite the global increase in anti-Asian racism due to the COVID-19 pandemic [12]. [14] Notice also that the images depicting Black individuals received the least amount of Race-related tags, as compared to other races, in 2018.

Given the difference in race-marking in our data, we could say that some workers might suffer from a form of implicit racism in the 2020 study. We have analyzed the number of inflammatory tags referring to people of color, and noticed a decrease in 2020 in comparison with the data of 2018. However, race-marking has clearly *increased* in 2020. This could be an indication that workers are thinking about what is "politically correct" when describing people, but involuntarily, they are still showing markedly different behavior towards people of color. Another factor that can be impacting our observed results, is that the race labels on each subject in an image is self-reported by the subject [25].

Finally, considering the Nationality tags, we observe a statistically significant increase in 2020, as compared to 2018 ($z = -4.09$, $p < .001$). This observation provides us with a ***partially positive*** answer this time regarding the *Identity* aspect of **RQ1**. Examining each group of images (by racial group), we observe more tags describing Nationality in 2020, as compared to 2018. However, the differences are not statistically significant, in particular, for the images depicting Black subjects ($z = -0.87$, $p > .05$).

Considering the sub-clusters of tags in the Nationality cluster, we noticed an increase in the number of occurrences in all of them. Noticeably, tags describing the depicted person as "American" or as a "USA citizen" show a statistically significant increase in use in 2020 ($z = -2.86$, $p = .004$). Factors that could have affected these results include the Black Lives Matter movement, the upcoming U.S. presidential elections, or the aforementioned anti-Chinese xenophobic sentiments due to the COVID-19 pandemic, all of which may be influencing the workers' "state of mind."

[14] www.hrw.org/news/2020/05/12/covid-19-fueling-anti-asian-racism-and-xenophobia-worldwide

5.2 Stylistic perspective: Inferential and Concrete language

The analysis presented above took a thematic perspective when considering the use of the *Health* and *Identity*-related tags. Next, we provide a stylistic perspective on the workers' descriptions of the depicted person, at the same time differentiating between the *Health* and *Identity* themes. Before moving on to look at the stylistic differences by theme, we examine whether the number of occurrences of the tags workers provided (i.e., Inferential, Concrete, or Other remaining tags) are independent of the year these tags were collected, which is our null hypothesis. We observe that the distributions of the tags by stylistic perspective differ between 2018 and 2020 ($X^2(2, N = 34923) = 10.64$, $p = .004$).

As we can see in Figure 5, workers provided significantly more inferential tags for the *Identity* theme in 2020 overall, as compared to 2018 ($z = -3.06$, $p = .002$). However, considering the images by racial group of the depicted persons, we find that workers became more inferential in 2020 only when describing the images of Black ($z = -5.33$, $p < .001$) and Asian ($z = -2.64$, $p = .008$) subjects. Instead, it appears that the proportions of the inferential tag occurrences between the two years are similar for images of White ($z = 2.12$, $p = .033$) and Latino/a ($z = 1.30$, $p > .05$) subjects considering a significance level of $p < .0125$. We consider the significance level for the four hypothesis above to be $p < .0125$ according to the Bonferroni correction.This is particularly interesting given the racial climate in the U.S. during the time frame that the 2020 data were collected, and seems to suggest a sensitivity towards the issue of identity and racial minorities. Frequency of use for the concrete tags is stable between 2018 and 2020, in the *Identity* theme.

Considering now the *Heath-related* theme, we can see an opposite trend. Overall, the concrete tags belonging to this theme are significantly more frequent in 2020 compared to 2018 ($z = -3.54$, $p < .001$), while inferential tags are overall stable across the two years. An interesting remark is that images of Black subjects are the only group receiving significantly more concrete *Health-related* tags in 2020 as compared to 2018 ($z = -5.09$, $p < .001$).

In answering **RQ2**, we can say that it appears that workers' use of *Identity*-related tags exhibited a notable change from 2018 to 2020. Overall, workers used more abstract tags in 2020, in particular, when describing images depicting Asian and Black individuals. In other words, workers made more inferences and assumptions about the depicted individual, instead of simply describing their physical appearance. This may imply that in 2020, workers had a heightened awareness of, or sensitivity towards, the issue of race. Notice that this is only one factor that can have impacted our results. As we discuss in Section 6.2, since we do not have control over the set of workers recruited in 2020, it is possible that we ended up with a large concentration of workers that are sensitive towards the issue of race, because of the particular time we have initiated the crowdsourcing task. In other words, there is no way for us to gauge if we have received an over-representation of responses from people in a particular region of the U.S. where racial tensions had intensified during this time. Given the large available pool of workers in the Appen platform, this a more unlikely factor, but might still have some effect on the observed results.

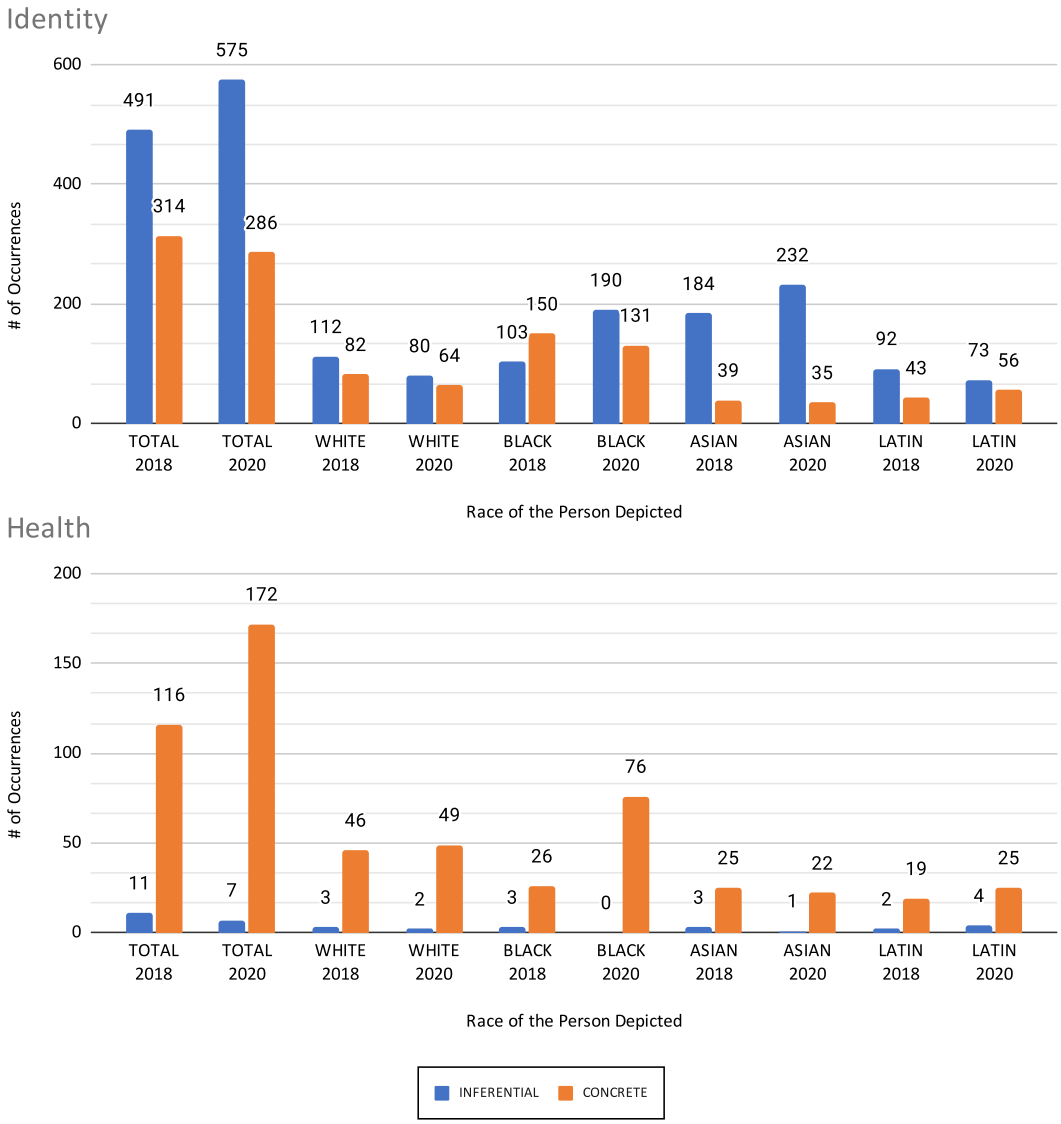

Figure 5: Stylistic analysis (inferential (left bar) v. concrete (right bar)) on *Identity* and *Health* tags, used across time, by racial group of depicted person.

6 DISCUSSION

Data collection through crowdsourcing, and in particular, for capturing the implicit, visual information contained in images, can be challenging, especially in the hands of non-experts, who may be unfamiliar with issues of data bias. Computer vision experts deal with the issue by mitigating bias in image datasets, during the process of algorithm development (e.g., [21, 47]). Bias mitigation during the collection or processing of the crowdsourcing results, is a field that is currently being explored (e.g., [1, 31]). Recently, due to the COVID-19 pandemic, crowdsourcing platforms are becoming busier, both in terms of supply (i.e., available workforce) as well as demand (e.g., with physical laboratories moving online). Moving on-premises laboratory studies to crowdsourcing platforms is not trivial, and thus, has been researched in the past [13] in terms of quality and reproducibility of the resulting data. What is presently unaccounted for is the potential influence that major local or global events can have on the collected data through crowdsourcing, and how this might affect the human bias introduced in datasets built over a large time frame (e.g., ImageNet).

Our findings support the existence of a parameter conforming with societal events, which alongside other factors, can influence and create variation in the data harvested through crowdsourcing. Thus, given a particular crowdsourcing task, it is essential that additional steps are taken, to identify whether the collected data

can be influenced by variations associated with societal events, and if so, whether measures need to be taken to restrain or promote said effects.

In this work, we looked at the consequential example of an open-ended annotation task on people-related images. To observe in a quantitative manner the impact of a temporal variation, we considered the number of unique occurrences as an indication of the popularity of a given tag of interest, among the workers. It is an indication of how frequently we will encounter a certain word in a similar image tagging setting. In the example of an automated process receiving the crowdsourced data for training a computer vision algorithm, it is more likely that the most popular tags (in our scenario, the tags with the highest number of occurrences) will be the ones to be incorporated and used. Applications trained on our 2018 versus 2020 datasets might be positively or negatively impacted, depending on the application's particular context. For instance, an image tagging algorithm trained on the 2020 data, might end up using more weight-related tags to describe input images of people, which could be seen as a risk when deployed in user-facing apps or platforms.

Descriptive image annotation is one type of task that can be impacted by temporal variations. Another example is image content moderation. In such a setting, data collected through crowdsourcing can assess the content of an image in an open-ended way (as we explored in this study), or in a closed form way, by providing a list of possible answers or a scale to measure, for instance, the degree of violence in the image. In this example, an image that in the past was considered inappropriate by the majority of crowdworkers, might not be in the present, and vice versa. Hence, the requester of such a dataset must be aware of this fact. It must be noted that image annotation is merely one class of crowdsourced micro-tasks that can have serious ramifications as a result of temporal variations. Future work can explore the types of temporal variations that may manifest in other common micro-tasks such as descriptive annotation on other media (e.g., audio or video), image categorization, or sentiment analysis.

6.1 Coping with temporal variations

This work has demonstrated that temporal variations can appear in the crowdsourced data and has discussed how these variations can generate, or even reduce, pre-existing social bias in a dataset. Given the crowdsourcing task we considered in this study, we noticed that the data collected in 2018 and 2020 had tags that varied in the number of occurrences on the topic of a person's identity and body weight. As we mention in Section 5, the major events of the 2020 pandemic and the BLM movement, are significant social factors that, among others, bring about this observed variation in the data. Below, we discuss the implications of our work, providing a set of instructions, using as a running example the crowdsourced datasets generated in this work. These instructions can be used as guidelines for managing the possible temporal variations that manifest when requesters use crowdsourcing for the purpose of collecting and/or annotating image datasets, regardless of whether the variations are considered desirable or not.

6.1.1 Awareness. As a first step, the requester of a crowdsourcing task must first identify the "nature" of the crowdsourcing task towards the generation of the dataset. A crowdsourcing task may involve generating the dataset (from the point of image collection) and/or enriching it. In this study, our task focused on enriching an existing dataset, by providing descriptive tags for the provided images. It is important for a requester to identify the nature of the task since generating an entirely new dataset implies that a complete record of the major global or local events taking place at the particular time and space, where the crowdworkers are "positioned," must be recorded. Recording this type of contextual information must be an essential part of the dataset, which will permit the requester or anyone else that will use the dataset, to be mindful of the reporting bias that it might potentially carry, or at a later stage when more data will be generated. On the other hand, when the crowdsourcing task focuses on enriching a pre-existing dataset, the requester must be aware of any bias pre-existing in the dataset, and take it into consideration when looking into temporal variations. In our case, the "ground truth" race label on each image was provided by the depicted people, who self-reported their race [25]. Thus, any conclusion we make regarding the observed temporal variation must take this fact into account, as was discussed in the analysis.

Besides identifying the nature of the crowdsourcing task, and being aware of the possible implications that this might have for the data's susceptibility to temporal variations, there is more that can be done to raise awareness. In the present study of an open-ended, annotation task on people images, we observed evidence that stressful social circumstances are reflected in the data collected. Thus, we would recommend requesters acquiring additional information from workers, such as: (1) the worker's familiarity with the content of the image, (2) whether the content of the image produces stress in any way to the worker, (3) whether the content of the image produces a positive or a negative feeling to the worker. Additionally, given the correlation we observed among the societal events and the reported dataset, the worker might be asked to identify the degree and intensity of information they receive on major societal events happening around the globe and in their present location. The above questions regarding the worker's possible "connection" to the image can be a valuable tool for the requester for identifying whether the collected data can be affected by temporal variations. Moreover, depending on the order in which these questions will be posed to the worker, they can also act as a tool for managing temporal variations (see Section 6.1.3).

6.1.2 Recognition. A second step that the requester must take is to acknowledge whether similar datasets have been created in the past, or whether there is an intention of creating similar datasets in the future. Recognizing this fact is essential since temporal variations will obviously have a greater impact on reproduced or enhanced datasets. At this step, the requester of the crowdsourcing task must be able to identify whether the information collected in the "Awareness" step indicates that collecting new crowdsourced data will reduce or augment the potential bias introduced in the datasets due to major events, and how this aligns with the use of the collected data. For example, if the data presented in this study were to be used to predict the body weight of a person, the 2020 data might prove to be more valuable, as compared to the 2018 dataset.[15] Thus, the requester of the data must be able to recognize whether the

[15]Note that this is simply an example, the data collected in this study were not collected with the goal of being used in weight prediction algorithms, and we would advise

collected data will produce a desirable or undesirable temporal variation according to their usage. This step is essential for understanding how to manage temporal variation, in particular if it is undesirable.

6.1.3 Management. Once the requester is aware that the data collected through crowdsourcing might be subject to undesirable temporal variations, a few steps can be taken to reduce said effects. A preventive step can be taken by posing the questions discussed in Section 6.1.1 before introducing the main crowdsourcing task. This action can raise the awareness of the crowdworkers towards the task; essentially, the workers are being pre-disposed towards the task [28]. We believe that this step may have a positive effect towards reducing temporal variations. An additional or alternative step, would be to prime workers by informing them of the goal of the task and the known inherent biases in the presented images. On the other hand, if the requester has identified that the potential temporal variation introduced during the crowdsourcing process is desirable, it is best to ask the questions introduced in the awareness step after the completion of the main task.

Each application will use the crowdsourced data according to its specific goal(s). It is possible that temporal variations affecting the collected data do not align with the goals of the application. For example, an algorithmic application predicting the most distinguishable features of a person, with the purpose of facilitating "search and find," must include universally accepted labels. Hence, our 2020 dataset – in which issues of Identity and Health are salient – must be used with additional caution in such an application. Mitigating bias in the collected data, once those data are collected, is an action strictly linked to the goal of the application. It appears that if data are to be used by a different application, it is a essential that the raw crowdsourced data are released, if possible. As part of the documentation, a section describing the specific time/main events during which the data were created, will help facilitate the re-use of the data. Documenting the procedures and the time-frame during which the data were created will help the user of these data to apply a similar logic to the "Recognition" step and aid in coping with temporal variations.

6.2 Limitations and Future Work

Due to the online nature of the study, and the manner in which crowdworkers are recruited, it was not possible for us to perform a controlled experiment; inviting the same workers to participate was naturally impossible. Thus, no claims can be made towards a strict causal relationship between the situation/time and the changes in workers' annotations. Additionally, our study included a single task. Nonetheless, it is logical that other types of open-ended tasks, and particularly those requiring workers to provide natural language answers, will be impacted in a similar matter. We plan to identify other open-ended crowdsourcing tasks that have been performed in the recent past and replicate them.

In this study, we isolated the tags relevant to a person's physical appearance on the themes of *Health* and *Identity*. Thus, we have not yet examined many tags that can potentially allow us to observe further the influence of the present social events on the data. In particular, mapping the workers' emotional state through the reported tags is of interest. Thus, future work can consider the valance of emotions expressed (if at all) when describing the CFD images.

Additionally, given our current understanding of the influence of social events on crowdsourced data, we plan to repeat the image annotation task, including a questionnaire like the one discussed above in Section 6, and measure the extent to which the self-reported questions can capture, or help explain, the temporal variations. Finally, temporal effects are only one factor affecting a worker's mood or attention, another can be deliberately pre-disposing a worker with a qualification test or audio-visual stimulus [1], or a questionnaire as discussed above. In the future, we also plan to further explore this direction.

7 CONCLUSIONS

Our work contributes by pointing out another limitation in the use of crowdsourced data. The empirical evidence of our study emphasizes the influence that significant events – in this case, in the sphere of public health and racial discrimination – can have on crowdsourced data, among other factors, in an image annotation task. We linked our work to the yet largely unexplored area of crowdworkers' mood and variations in the collected data, by exploring the thematic and stylistic changes in workers' reports over time. Furthermore, our work extends the discussion concerning the repeatability of a crowdsourcing task and the replication of data. Although a task executed in the same platform can weakly conform to the repeatability condition, as pointed out in [35], we observe that replication of the results might not be feasible over periods of time marked by significant, large-scale events and/or experiences. We provide some interesting observations regarding the human crowdworkers' behavior that can be linked to attentional bias and race-marking, which is worth exploring further. The computer vision community, which relies extensively on human-labelled image datasets, now has to face a new crowdsourcing challenge especially during this sensitive period of time. To this respect, we provide a set of guidelines for recognizing and coping with temporal effects when requesting or using crowdsourced image datasets.

ACKNOWLEDGMENTS

This project is partially funded by the Cyprus Research and Innovation Foundation under grant EXCELLENCE/0918/0086 (DESCANT) and by the European Union's Horizon 2020 Research and Innovation Programme under agreements No. 739578 (RISE) and 810105 (CyCAT).

We would like to thank the anonymous reviewers for their valuable comments and insightful feedback and in particular our shepherd for support all along the way.

REFERENCES

[1] Natã M. Barbosa and Monchu Chen. 2019. Rehumanized Crowdsourcing: A Labeling Framework Addressing Bias and Ethics in Machine Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300773

[2] Pınar Barlas, Kyriakos Kyriakou, Styliani Kleanthous, and Jahna Otterbacher. 2019. Social B(eye)as: Human and Machine Descriptions of People Images. *Proceedings*

against it. This study advocates that crowdsourced data must be used almost exclusively for the purpose they were created for.

of the International AAAI Conference on Web and Social Media 13, 01 (Jul. 2019), 583–591. https://ojs.aaai.org/index.php/ICWSM/article/view/3255

[3] Alexander C Berg, Tamara L Berg, Hal Daume, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, and Kota Yamaguchi. 2012. Understanding and predicting importance in images. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Providence, RI, 3562–3569. https://doi.org/10.1109/CVPR.2012.6248100

[4] Camiel J Beukeboom et al. 2014. Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. *Social cognition and communication* 31 (2014), 313–330.

[5] Surabhi Bhutani and Jamie A. Cooper. 2020. COVID-19–Related Home Confinement in Adults: Weight Gain Risks and Opportunities. *Obesity* 28, 9 (2020), 1576–1577. https://doi.org/10.1002/oby.22904 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/oby.22904

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *CoRR* abs/1504.00325 (2015). arXiv:1504.00325 http://arxiv.org/abs/1504.00325

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

[9] James M Clay and Matthew O Parker. 2020. Alcohol use and misuse during the COVID-19 pandemic: a potential public health crisis? *The Lancet Public Health* 5, 5 (2020), e259.

[10] Michael A Cohn, Matthias R Mehl, and James W Pennebaker. 2004. Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science* 15, 10 (2004), 687–693.

[11] Lyn Craig and Brendan Churchill. 2020. Dual-earner parent couples' work and care during COVID-19. *Gender, Work & Organization* ., . (2020). https://doi.org/10.1111/gwao.12497

[12] Stephen M Croucher, Thao Nguyen, and Diyako Rahmani. 2020. Prejudice Toward Asian Americans in the Covid-19 Pandemic: The Effects of Social Media Use in the United States. *Frontiers in Communication* 5 (2020), 39.

[13] Matthew JC Crump, John V McDonnell, and Todd M Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one* 8, 3 (2013), e57410.

[14] Dhaval M Dave, Andrew I Friedson, Kyutaro Matsuzawa, Joseph J Sabia, and Samuel Safford. 2020. *Black Lives Matter protests, social distancing, and COVID-19*. Technical Report. National Bureau of Economic Research.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Miami, FL, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[16] Sylvie Droit-Volet, Sandrine Gil, Natalia Martinelli, Nicolas Andant, Maélys Clinchamps, Lénise Parreira, Karine Rouffiac, Michael Dambrun, Pascal Huguet, Benoît Dubuis, et al. 2020. Time and Covid-19 stress in the lockdown situation: Time free,«Dying» of boredom and sadness. *PloS one* 15, 8 (2020), e0236465.

[17] Anca Dumitrache. 2015. Crowdsourcing Disagreement for Collecting Semantic Annotation. In *The Semantic Web. Latest Advances and New Domains*, Fabien Gandon, Marta Sabou, Harald Sack, Claudia d'Amato, Philippe Cudré-Mauroux, and Antoine Zimmermann (Eds.). Springer International Publishing, Cham, 701–710.

[18] Itziar Fernández, Darío Páez, and James W Pennebaker. 2009. Comparison of expressive writing after the terrorist attacks of September 11th and March 11th. *International journal of clinical and health psychology* 9, 1 (2009), 89–103.

[19] Fernando Fernández-Aranda, Miquel Casas, Laurence Claes, Danielle Clark Bryan, Angela Favaro, Roser Granero, Carlota Gudiol, Susana Jiménez-Murcia, Andreas Karwautz, Daniel Le Grange, et al. 2020. COVID-19 and implications for eating disorders. *European Eating Disorders Review* 28, 3 (2020), 239.

[20] Susan T Fiske and Martha G Cox. 1979. Person concepts: The effect of target familiarity and descriptive purpose on the process of describing others 1. *Journal of Personality* 47, 1 (1979), 136–161.

[21] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women Also Snowboard: Overcoming Bias in Captioning Models. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 793–811.

[22] Kyriakos Kyriakou, Pınar Barlas, Styliani Kleanthous, and Jahna Otterbacher. 2019. Fairness in Proprietary Image Tagging Algorithms: A Cross-Platform Audit on People Images. *Proceedings of the International AAAI Conference on Web and Social Media* 13, 01 (Jul. 2019), 313–322. https://ojs.aaai.org/index.php/ICWSM/article/view/3232

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755.

[24] Stella F. Lourenco and Arber Tasimi. 2020. No Participant Left Behind: Conducting Science During COVID-19. *Trends in Cognitive Sciences* 24, 8 (2020), 583 – 584. https://doi.org/10.1016/j.tics.2020.05.003

[25] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods* 47, 4 (2015), 1122–1135.

[26] Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing Through the Human Reporting Bias: Visual Classifiers From Noisy Human-Centric Labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, 2930–2939.

[27] Robert R Morris, Mira Dontcheva, Adam Finkelstein, and Elizabeth Gerber. 2013. Affect and Creative Performance on Crowdsourcing Platforms. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, Geneva, 67–72. https://doi.org/10.1109/ACII.2013.18

[28] Robert R Morris, Mira Dontcheva, and Elizabeth M Gerber. 2012. Priming for better performance in microtask crowdsourcing environments. *IEEE Internet Computing* 16, 5 (2012), 13–19.

[29] Laura R Novick. 2003. At the forefront of thought: The effect of media exposure on airplane typicality. *Psychonomic bulletin & review* 10, 4 (2003), 971–974.

[30] Jahna Otterbacher. 2015. Crowdsourcing Stereotypes: Linguistic Bias in Metadata Generated via GWAP. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1955–1964. https://doi.org/10.1145/2702123.2702151

[31] Jahna Otterbacher. 2018. Social Cues, Social Biases: Stereotypes in Annotations on People Images. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 6, 1 (Jun. 2018). https://ojs.aaai.org/index.php/HCOMP/article/view/13320

[32] Jahna Otterbacher, Pınar Barlas, Styliani Kleanthous, and Kyriakos Kyriakou. 2019. How Do We Talk about Other People? Group (Un) Fairness in Natural Language Image Descriptions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 106–114. https://ojs.aaai.org/index.php/HCOMP/article/view/5267

[33] Crystal L Park, Beth S Russell, Michael Fendrich, Lucy Finkelstein-Fox, Morica Hutchison, and Jessica Becker. 2020. Americans' COVID-19 Stress, Coping, and Adherence to CDC Guidelines. *Journal of General Internal Medicine* 8, 35 (2020), 2296–2303.

[34] Betty Pfefferbaum and Carol S. North. 2020. Mental Health and the Covid-19 Pandemic. *New England Journal of Medicine* 383, 6 (2020), 510–512. https://doi.org/10.1056/NEJMp2008017 PMID: 32283003.

[35] Rehab Qarout, Alessandro Checco, Gianluca Demartini, and Kalina Bontcheva. 2019. Platform-Related Factors in Repeatability and Reproducibility of Crowdsourcing Tasks. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 135–143. https://ojs.aaai.org/index.php/HCOMP/article/view/5264

[36] Marian Sauter, Dejan Draschkow, and Wolfgang Mack. 2020. Building, Hosting and Recruiting: A Brief Introduction to Running Behavioral Experiments Online. *Brain Sciences* 10, 4 (2020), 251.

[37] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.

[38] Gün R Semin and Klaus Fiedler. 1988. The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of personality and Social Psychology* 54, 4 (1988), 558.

[39] Ari Shechter, Franchesca Diaz, Nathalie Moise, D Edmund Anstey, Siqin Ye, Sachin Agarwal, Jeffrey L Birk, Daniel Brodie, Diane E Cannone, Bernard Chang, et al. 2020. Psychological distress, coping behaviors, and preferences for support among New York healthcare workers during the COVID-19 pandemic. *General hospital psychiatry* 66 (2020), 1–8.

[40] Han Yu Zhiqi Shen, Simon Fauvel, and Lizhen Cui. 2017. Efficient scheduling in crowdsourcing based on workers' mood. In *2017 IEEE International Conference on Agents (ICA)*. IEEE, Beijing, 121–126.

[41] Leo Sher. 2020. COVID-19, anxiety, sleep disturbances and suicide. *Sleep Medicine* 70, 124 (2020).

[42] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[43] Jet D. Termorshuizen, Hunna J. Watson, Laura M. Thornton, Stina Borg, Rachael E. Flatt, Casey M. MacDermod, Lauren E. Harper, Eric F. van Furth, Christine M. Peat, and Cynthia M. Bulik. 2020. Early impact of COVID-19 on individuals with self-reported eating disorders: A survey

of 1,000 individuals in the United States and the Netherlands. *International Journal of Eating Disorders* ., . (2020). https://doi.org/10.1002/eat.23353 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/eat.23353

[44] Emiel van Miltenburg. 2016. Stereotyping and Bias in the Flickr30K Dataset. *CoRR* abs/1605.06083 (2016). arXiv:1605.06083 http://arxiv.org/abs/1605.06083

[45] Natalie N Watson-Singleton, Yara Mekawi, Kaleigh V Wilkins, and Isatou F. Jatta. 2020. Racism's effect on depressive symptoms: Examining perseverative cognition and Black Lives Matter activism as moderators. *Journal of Counseling Psychology* Advance online publication. (2020). https://doi.org/10.1037/cou0000436

[46] Janine Willis and Alexander Todorov. 2006. First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological science* 17, 7 (2006), 592–598.

[47] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2979–2989. https://doi.org/10.18653/v1/D17-1323

[48] Mengdie Zhuang and Ujwal Gadiraju. 2019. In What Mood Are You Today? An Analysis of Crowd Workers' Mood, Performance and Engagement. In *Proceedings of the 10th ACM Conference on Web Science* (Boston, Massachusetts, USA) *(WebSci '19)*. Association for Computing Machinery, New York, NY, USA, 373–382. https://doi.org/10.1145/3292522.3326010