Crowdwork as a Snapshot in Time: Image Annotation Tasks during a Pandemic

Evgenia Christoforou, [1] **Pınar Barlas,**[1] **Jahna Otterbacher**[1][2]
[1]Research Centre on Interactive Media, Smart Systems and Emerging Technologies
[2]Open University of Cyprus, Nicosia, CYPRUS
{e.christoforou, p.barlas, j.otterbacher}@rise.org.cy

Abstract

While crowdsourcing provides a convenient solution for tapping into human intelligence, a concern is the bias inherent in the data collected. Events related to the COVID-19 pandemic had an impact on people globally, and crowdworkers were no exception. Given the evidence concerning mood and stress on work, we explore how *temporal events* might affect crowdsourced data. We replicated an image annotation task conducted in 2018, in which workers describe people images. We expected 2020 annotations to contain more references to health, as compared to 2018 data. Overall, we find no evidence that health-related tags were used more often in 2020, but instead we find a significant increase in the use of tags related to *weight* (e.g., fat, chubby, overweight). This result, coupled with the "stay at home" act in effect in 2020, illustrate how crowdwork is impacted by temporal events.

The COVID-19 pandemic is proving to be a public emergency without precedent. Worldwide, individuals are feeling the impact; e.g., they may be suffering from confusion, isolation, and feelings of insecurity (Pfefferbaum and North 2020). Public health researchers are citing large-scale problems such as alcohol and drug abuse (Clay and Parker 2020) as well as increased levels of anxiety and sleep disturbances (e.g., insomnia) (Sher 2020). Other concerns include increased issues with eating disorders (Fernández-Aranda et al. 2020) and the danger of weight gains in individuals over short periods of time (Bhutani and Cooper 2020).

The "stay at home" movement and/or enforced lockdowns, in combination with the economic crisis during 2020, have created a fruitful environment for crowdwork supply to bloom and demand to increase[1], with on-premise laboratory studies being suspended in many areas. A plethora of new datasets produced through crowdsourcing are being created; but can crowdwork during a pandemic yield reliable data? We seize the opportunity to observe crowdworkers during what clearly cannot be considered "normal times," and provide evidence that *temporal events* can introduce "noise" in crowdsourced data.

Contribution. We replicated an image annotation task conducted in 2018 on a crowdsourcing platform. The task prompted each worker to provide 10 tags that describe the image content. All images depicted one person in a "passport-style" photo. We *hypothesize* that, in the replicated study of 2020, we would observe an increase in the vocabulary size and frequency of health-related tags. Contrary to our expectations, we found no evidence supporting our arguments. Surprisingly, we observed a significant increase in the vocabulary and frequency of tags describing a person's excess weight. Our data suggest that "staying at home" introduced an unexpected noise parameter in the resulting annotations produced by crowdworkers.

Related Work. Lourenco and Tasimi raise the question of how running crowd studies during the pandemic (and considering as well the stay-at-home orders in place) affects the distribution of participants (Lourenco and Tasimi 2020). Inspired by findings on the physical workplace, (Zhuang and Gadiraju 2019) surveyed crowdworkers to understand how a worker's mood affects his or her work. They studied both the worker's perception of how moods impact work, as well as their actual performance on an information-finding task. Although the findings showed that workers' moods affected their engagement with a task, that was not reflected in the actual evaluation of the work.

Methodology

We replicated the methodology used in the 2018 image annotation task (Barlas et al. 2019). Our crowdsourcing study in both years was powered by the Appen[2] platform. We restricted our participants to the U.S. in both years and asked for three unique crowdworker judgements per image. The task was composed of 597 images in total: 183 depicted White, 197 depicted Black, 109 depicted Asian and 108 Latino/a persons. Each worker was permitted to judge up to 20 images and was required to provide 10 tags per image. In 2020, our study was active from May to June, while the previous task was active December 2018, 18 months apart.

The collected tags were pre-processed by applying the same spell-check process as in 2018 and translating the Spanish-language tags collected in both years into English. In the 2018 study, all processed tags were grouped into clusters following a specific typology (see (Barlas et al. 2019), Table 4). For the needs of this study, we processed the col-

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

[2]https://appen.com/. Experiments in 2018 were executed using the FigureEight platform, later acquired by Appen.

	Grouped Tags	2019	2020	p-value
Health	**Symptoms:** {sick, runny_nose, hurt, pain, pained, pained_expression, ache, some_ache, dehydrated, sweaty, sweat, sweating, perspiring, sleeping_problems, bloody, dilated_pupils, bloodshot_eyes, sick_eye, eye_problem, eye_bubble}	16	18	0.7304
	Overall health: {health, healthy, healthy_glow, unhealthy, fit, cadaverous}	11	7	–
Weight	**Overweight:** {chubby, chubby_girl, chunky, fat, fat_boy, fat_feminine, fat_girl, fat_guy, fat_man, fat_masculine, fat_person, fat_woman, fat_women, fatty, heavy, obese, overweight, corpulent, plump, full_figure}	63	103	0.0015
	Underweight: {skeletal, bony, skinny, skinny_boy, skinny_girl, skinny_woman, slender, slim, thin, thin_man, thin_person, thin_woman}	33	51	0.0469
	Normal weight: {normal_weight, average_weight}	4	0	–

Table 1: Categorization into sub-clusters of the unique tags for the health and weight clusters. Third and fourth column depicts the total number of unique occurrences for each sub-cluster in the two years of the study. The last column depicts the p-value.

lected tags from both years identifying the tags broadly related to *health*. Our initial processing of the data indicated that a set of tags referring to body weight had a larger cardinally than tags referring directly to health conditions. For this reason, our analysis considers two main clusters: (1) health and (2) weight. The health cluster consists of two sub-clusters: symptoms and overall health. The symptoms sub-cluster contains all the tags that could be used to describe a health condition while the overall health sub-cluster contains all the tags that describe whether a person is healthy or not. Tags referring to a physiological characteristic of a person that might be related with a health condition (i.e., "albino", "broken-nosed", etc.) were not included. Furthermore, ambiguous tags referring to the color of a body part (i.e., "pale_face") were also excluded as being too ambiguous. The weight cluster consists of three sub-clusters: overweight, underweight and normal weight. Tags describing the person's body structure (i.e., "heavy_build") were excluded.

Considering the above categorization, two researchers went through the entire set of unique tags collected in both years, placing each tag either in one of the sub-clusters or discarding it. Their results were compared and the disagreements were discussed and resolved by a third researcher.

Results

For each tag belonging to a sub-cluster, we computed the number of unique occurrences in each of the two studies. 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. Table 1 provides the aggregated number of unique occurrences for tags used in 2018 and 2020 and the p-value of a two proportion z-test, considering 1.791 possible unique occurrences of a tag over the total number of judgments.

Our null hypothesis considers that results in 2018 and 2020 are equal when the p-value is > 0.05. Thus, significantly more tags focused on depicted persons being *overweight* in 2020, in the midst of the pandemic, as compared to the 2018 study. However, the results concerning the *underweight* sub-cluster are rather inconclusive since the p-value is close to the threshold.

Additionally, our analysis showed that the set of images receiving an "overweight" tag had an almost constant ratio of female (f) and male (m) subjects depicted in 2018 (0.6f:0.4m) and 2020 (0.65f:0.35m). Moreover, workers in 2018 and 2020 agreed on tagging the same 27 images with an "overweight" tag, while in 2020, we had 50 distinct images versus 16 in 2018 receiving an "overweight" tag. Regarding the images receiving an "underweight" tag we had the majority of images shifting from male in 2018 (0.4f:0.6m) to female in 2020 (0.64f:0.36m). It was interesting to notice that workers in 2018 and 2020 only agreed on tagging two images with the "underweight" tag, while they selected 40 distinct images in 2020 versus 28 in 2018. Finally, it appears that workers in 2020 did not use health-related tags more frequently to describe the person images. In fact, even if not significantly larger, we observed more occurrences in the overall health sub-cluster in 2018.

Discussion and Future Work

The images used in the task were from the Chicago Face Database (Ma, Correll, and Wittenbrink 2015) and depict healthy individuals in a passport-style photo. Thus, the lack of large numbers of health-related tags during the pandemic testifies in favor of the crowdworkers as a valid source of human-intelligence. Nonetheless, a significantly larger, more diverse set of tags was used to describe the excess weight of a person in 2020. It appears that workers developed a new sensitivity towards annotating images of people. It is possible that this is associated with attentional bias on the part of the workers. In fact, according to Google Trends [3], 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.

Regardless of the reasons behind the bias developed by crowdworkers, the fact remains that the 2020 pandemic introduced a "noise" in the tags related to weight, influencing our datasets. In conclusion, temporal events are clearly a factor affecting crowdsourcing quality and we plan to explore what other categories of tags in our human annotation study were impacted by it.

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), 810105 (CyCAT).

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

References

[Barlas et al. 2019] Barlas, P.; Kyriakou, K.; Kleanthous, S.; and Otterbacher, J. 2019. Social b(eye)as: Human and machine descriptions of people images. *Proceedings of the International AAAI Conference on Web and Social Media* 13(01):583–591.

[Bhutani and Cooper 2020] Bhutani, S., and Cooper, J. A. 2020. Covid-19 related home confinement in adults: weight gain risks and opportunities. *Obesity*.

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

[Fernández-Aranda et al. 2020] Fernández-Aranda, F.; Casas, M.; Claes, L.; Bryan, D. C.; Favaro, A.; Granero, R.; Gudiol, C.; Jiménez-Murcia, S.; Karwautz, A.; Le Grange, D.; et al. 2020. Covid-19 and implications for eating disorders. *European Eating Disorders Review* 28(3):239.

[Lourenco and Tasimi 2020] Lourenco, S. F., and Tasimi, A. 2020. No participant left behind: Conducting science during covid-19. *Trends in Cognitive Sciences*.

[Ma, Correll, and Wittenbrink 2015] Ma, D. S.; Correll, J.; and Wittenbrink, B. 2015. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods* 47(4):1122–1135.

[Pfefferbaum and North 2020] Pfefferbaum, B., and North, C. S. 2020. Mental health and the covid-19 pandemic. *New England Journal of Medicine*.

[Sher 2020] Sher, L. 2020. Covid-19, anxiety, sleep disturbances and suicide. *Sleep Medicine*.

[Zhuang and Gadiraju 2019] Zhuang, M., and Gadiraju, U. 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*, 373–382.