Paid Crowdsourcing, Low Income Contributors, and Subjectivity

Giannis Haralabopoulos, Christian Wagner, Derek Mcauley, Ioannis Anagnostopoulos

Paid crowdsourcing, low income contributors, and subjectivity

Giannis Haralabopoulos[1][0000−0002−2142−4975], Christian Wagner[1], Derek McAuley[1], and Ioannis Anagnostopoulos[2]

[1] University of Nottingham, UK
name.surname@nottingham.ac.uk
[2] University of Thessaly, Greece
janag@dib.uth.gr

Abstract. Scientific projects that require human computation often resort to crowdsourcing. Interested individuals can contribute to a crowdsourcing task, essentially contributing towards the project's goals. To motivate participation and engagement, scientists use a variety of reward mechanisms. The most common motivation, and the one that yields the fastest results, is monetary rewards. By using monetary, scientists address a wider audience to participate in the task. As the payment is below minimum wage for developed economies, users from developing countries are more eager to participate. In subjective tasks, or tasks that cannot be validated through a right or wrong type of validation, monetary incentives could contrast with the much needed quality of submissions. We perform a subjective crowdsourcing task, emotion annotation, and compare the quality of the answers from contributors of varying income levels, based on the Gross Domestic Product. The results indicate a different contribution process between contributors from varying GDP regions. Low income contributors, possibly driven by the monetary incentive, submit low quality answers at a higher pace, while high income contributors provide diverse answers at a slower pace.

Keywords: Crowdsourcing · Demographics · Monetary Rewards · Subjectivity.

1 Introduction

Crowdsourcing is the process where a number of non expert people collectively perform a task. The task is presented to a wide group of internet users via an online platform and each user provides their unique input. The collection of inputs from the crowd is aggregated to provide information necessary for the task at hand. The users are referred as workers or contributors and the person that requests the crowdsourcing is known as requester.

As more and more requesters create tasks for workers, new tasks must provide incentives for participation. Money is the most common crowdsourcing incentive, and the one that yields the fastest results, but of varying quality [1]. Gamification

of crowdsourcing tasks, i.e. the use of game elements such as point ladders and rewards, is another way of keeping workers interested and engaged in the task [8].

A crowdsourcing task is split in small micro tasks. Each micro task requires some seconds (on rare occasions minutes) to be completed, therefore the monetary incentive per micro task is usually low. The minimum allowed per micro task is 0.01$ in most crowdsourcing platforms[3][4]. This results in a low salary per hour, which according to Horton and Chilton [5], has a median of 1.38$/hour.

Studies have showed that workers from Europe and America provide higher quality contributions than workers from Asia [6,7]. In addition, it is also shown that per task payment reduces contributor productivity [13] and monetary rewards negatively influence answers quality [14]. Considering that Asia's mean GDP per capita is almost 26% of Europe's, less than 14% of North America's[5], and the fact that participation incentives exist due monetary rewards, could there be a link between income and quality of crowdsourcing contributions?

During the analysis of the crowdsourcing results from one of our recent studies [9,12], we found evidence of high percentage of spam or dishonest contribution from workers originating from low income countries. Participants that deliver systematically the same contributions were identified as dishonest contributors, with high certainty. We hypothesize that workers from low income countries are mainly motivated by the monetary incentive, rather than the contribution to the crowdsourcing task.

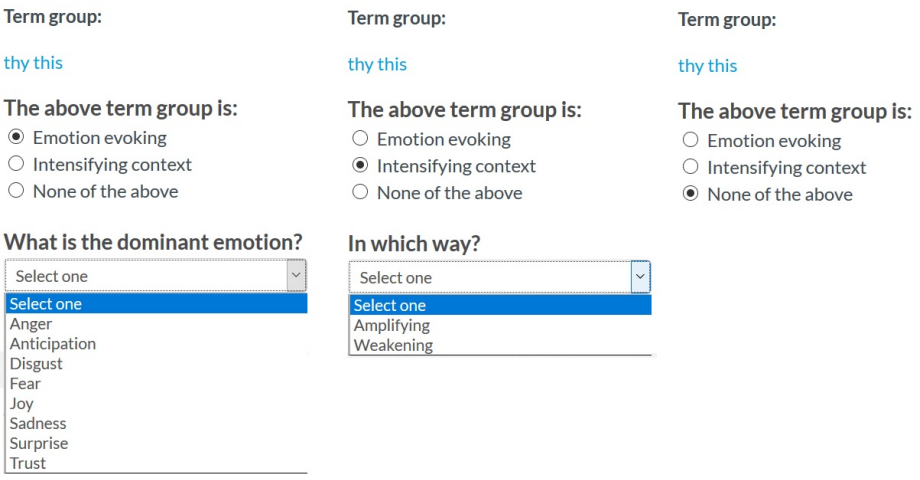

Fig. 1. Micro-task structure

[3] https://www.figure-eight.com/
[4] https://www.mturk.com/
[5] http://www.imf.org/external/datamapper/NGDPDPC@WEO/OEMDC/ADVEC/WEOWORLD

2 Task

The crowdsourcing task, as described in [9,12], requires workers to choose from one of the three main classes and then select the most appropriate subclass. As seen in Fig.1, workers select the appropriate class with a radio button, and unless the class selection is none, a drop-down menu appears that requires a single selection.

The use of fixed radio and drop-down selections, allows the requester to capture and filter out spamming behavior. This type of behavior is easy to capture when a dishonest worker is constantly selecting the same options over and over again, but undetectable if a dishonest user utilizes more elaborate methods of spamming, such as automated mouse movement applications. The inclusion of quality control questions [10], which usually appear at the start of the task, forces workers to adopt a cautious behavior early on the task, and then proceed to provide dishonest contributions at a higher pace. In addition, the nature of a subjective task, such as the emotional annotation of terms, doesn't provide a solid basis for quality control questions that use predefined answers.

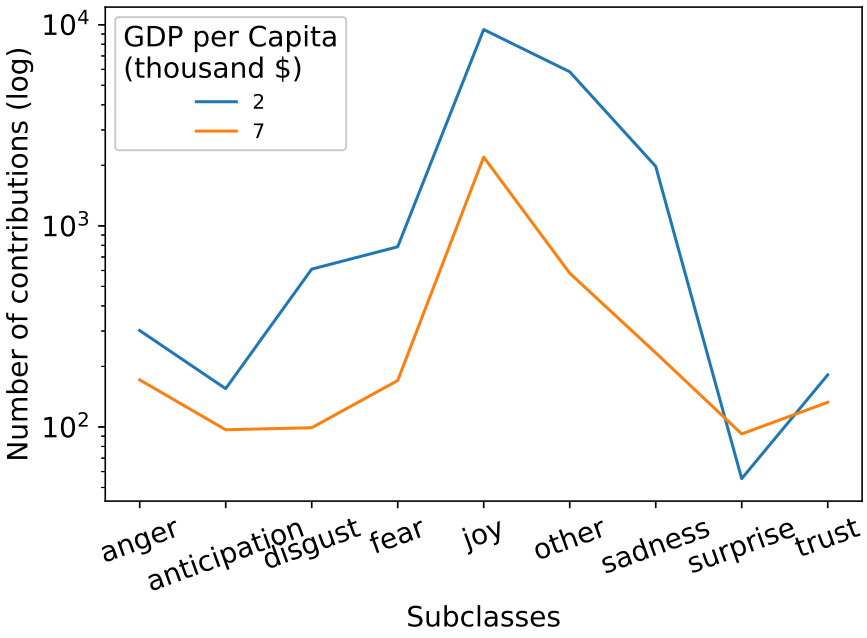

Fig. 2. Subclass distribution (low income levels)

In this subjective crowdsourcing task, each term group was annotated by 6 separate workers, and each worker was not able to annotate more than 1%

of the corpus. The subclass distribution of the contributions, as seen in Fig.2, was comprised of mostly joy annotated term groups. The nine subclasses, visible in Fig.2 and Fig.3, represent the eight emotions, while the subclass "other" refers to either intensifiers and negators, or term groups that are not emotion evocative and not intensifying. Ninety and seventy two per cent of workers from countries with GDP income per capita of less than 10K$ annotated term groups exclusively with joy or none. This high annotation percentage in a single subclass, of a random distribution of terms, is strong indication of spam and dishonesty.

Although this behavior was verified manually by the authors at a term level, where we analysed individual terms to determine whether they were evoking any emotion, in large scale crowdsourcing tasks there is no need for manual verification. The probability of annotation workers constantly encountering terms of the same emotion would be extremely low. E.g. if the combined probability of a term group being none or joy was 99%, the probability of encountering six hundred exclusively none or joy term groups would only be 0.2405%. Thus the probability that a worker honestly annotated a diverse corpus of words only with a single emotion is very close to zero.

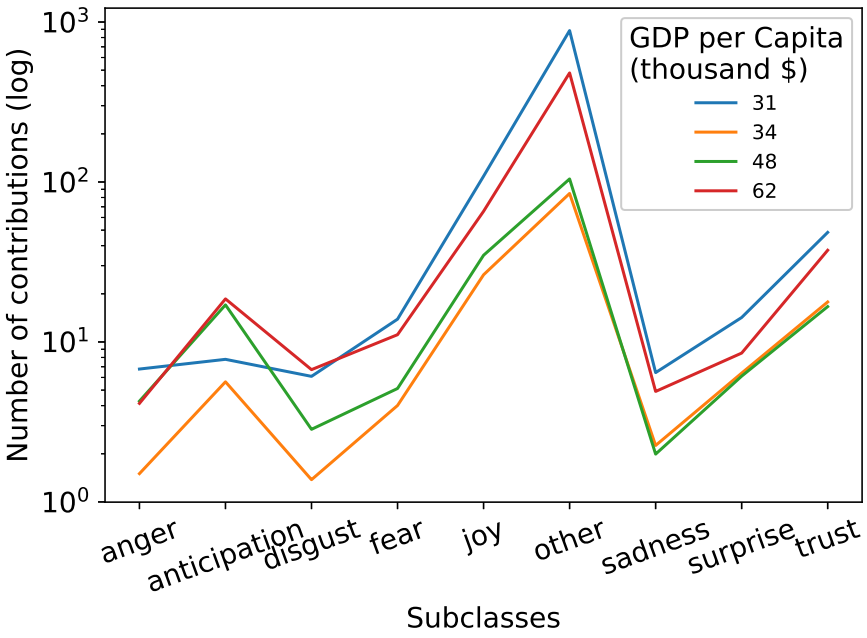

Fig. 3. Subclass distribution in previously joy or other annotated term groups (high income levels)

3 Quality of workers and contributions

The requested task had both a set of test questions, that would say if a term was an intensifier or emotion evoking, and a distribution control of contributions was applied. However, a high number of test questions would appear during the initial stages of the task, something that workers were aware of and could easily exploit. The distribution control worked well on identifying spammers, but it required a minimum number of answers from each worker before excluding workers that constantly annotated a single emotion.

The distribution threshold was set at 40%, which is approximately double the highest occurrence of a single emotion in another similar study [11]. After filtering out workers that had more than 40% of their total annotations in a single subclass, we end up with a smoother distribution of annotations, but strictly bound to the threshold. From the total one hundred eighty seven participating workers, only thirty six were identified as eligible based on the 40% filtering process. This in turn, reduced the number of total annotations deemed valid, from more than fifty thousands to less than eleven thousands. The percentage of invalidated annotations and the originating country of the contributors are the basis of our hypothesis.

4 Testing the Hypothesis

Our hypothesis is that workers from low income countries are motivated from the monetary incentive, rather than contributing to the task. A direct implication of this hypothesis is that monetary incentivised workers provide low quality answers. In order to test this hypothesis we used 1555 term groups as annotated strictly to joy or none, prior to the filtering process, with a majority higher than 80% of the total annotations. Six annotations per term group were required, with the same test questions for workers as before. The task was requested in the same platform, with the exact same settings, but only allowed workers from countries with GDP per capita higher then 30K$ to participate.

In the high income task, workers annotated term groups diversely, and the previously dominant joy emotion is the third most frequently occurring subclass. Additionally, strong majority agreement (over 80%) over joy and none subclasses is lower, with 42 term groups annotated as joy and 181 as none, compared to 1055 and 500 respective term groups of the initial task.

In total group of 112 contributors in the hypothesis group, 32 were flagged as eligible by our 40% single annotation check. A slightly lower eligibility over assessment than the initial task, due to the fact that hypothesis term groups were biased towards joy and none. However, workers annotated term groups in all of the possible subclasses, and only 4% and 36.2% of the term groups were still majorly annotated as joy and none respectively.

None was the most common occurring annotation post-filter, and the only subclass affected by the single annotation check. Its challenging to determine honesty post-annotation with an unsupervised method, more so on a monetary incentivised subjective task.

5 Conclusion

The results suggest that income is an important factor in crowdsourcing participation. As the particular crowdsourcing task is mainly of subjective nature, the validation of the hypothesis can only be based on the diversity of the provided contributions, compared to the initially single subclass accumulated contributions. In an objective task, quality questions and distribution filtering are sufficient measures to prevent dishonesty. Modern crowdsourcing platforms[6] provide ethical rewards and worker demographic pre-screening, empowering requesters and workers alike.

We are studying the subjective aspects of crowdsourcing, and our goal is to address the evaluation of subjective contributions with objective criteria in future publications. Additionally, in purely subjective crowdsourcing tasks, demographics like age, sex, or education levels, should be further studied to better understand crowds performance [2,3]. Our preliminary findings are in line with research that suggests voluntary crowdsourcing provides highest quality contributions than paid crowdsourcing[1]. Workers participating voluntarily are only motivated by their desire to contribute, which can be supplemented by a range of non monetary incentives to improve engagement. Apart from voluntary participation, a critical mass of honest participators [4] can function as a self-maintained filter for dishonesty and spamming in both objective and subjective tasks.

References

1. Mao, Andrew, Ece Kamar, Yiling Chen, Eric Horvitz, Megan E. Schwamb, Chris J. Lintott, and Arfon M. Smith. "Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing." In First AAAI conference on human computation and crowdsourcing. 2013.
2. Pavlick, Ellie, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. "The language demographics of amazon mechanical turk." Transactions of the Association for Computational Linguistics 2 (2014): 79-92.
3. Ross, Joel, Lilly Irani, M. Silberman, Andrew Zaldivar, and Bill Tomlinson. "Who are the crowdworkers?: shifting demographics in mechanical turk." In CHI'10 extended abstracts on Human factors in computing systems, pp. 2863-2872. ACM, 2010.
4. Sharma, Ankit. "Crowdsourcing Critical Success Factor Model: Strategies to harness the collective intelligence of the crowd." London School of Economics (LSE), London (2010).
5. Horton, John Joseph and Chilton Lydia B. "The labor economics of paid crowdsourcing." In Proceedings of the 11th ACM conference on Electronic commerce (2010): 209-218
6. Rogstadius, Jakob, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. "An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets." ICWSM 11 (2011): 17-21.

[6] https://prolific.ac/

7. Kazai, Gabriella, Jaap Kamps, and Natasa Milic-Frayling. "The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy." In Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 2583-2586. ACM, 2012.
8. Hamari, Juho, Jonna Koivisto, and Harri Sarsa. "Does gamification work?–a literature review of empirical studies on gamification." In System Sciences (HICSS), 2014 47th Hawaii International Conference on, pp. 3025-3034. IEEE, 2014.
9. Haralabopoulos, Giannis, and Elena Simperl. "Crowdsourcing for Beyond Polarity Sentiment Analysis A Pure Emotion Lexicon." arXiv preprint arXiv:1710.04203 (2017).
10. Allahbakhsh, Mohammad, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar. "Quality control in crowdsourcing systems: Issues and directions." IEEE Internet Computing 17, no. 2 (2013): 76-81.
11. Mohammad, Saif M., and Peter D. Turney. "Nrc emotion lexicon." NRC Technical Report (2013).
12. Haralabopoulos, G., Wagner, C., McAuley, D., and Simperl, E. (2018, October). A Multivalued Emotion Lexicon Created and Evaluated by the Crowd. In 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 355-362). IEEE
13. Ikeda, Kazushi, and Michael S. Bernstein. "Pay it backward: Per-task payments on crowdsourcing platforms reduce productivity." Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, 2016.
14. Acar, Oguz Ali. "Harnessing the creative potential of consumers: money, participation, and creativity in idea crowdsourcing." Marketing Letters 29.2 (2018): 177-188.