Investigating Stability and Reliability of Crowdsourcing Output

Rehab Kamal Qarout[1], Alessandro Checco[2], and Kalina Bontcheva[1]

[1] University of Sheffield, Department of Computer Science, Sheffield, UK
[2] University of Sheffield, Information School, Sheffield, UK
{rkqarout1, a.checco, k.bontcheva}@sheffield.ac.uk

Abstract. This research proposes to investigate the reliability of the output of crowdsourcing platforms and its consistency over time. We study the effect of design interface and instructions and identify critical differences between two platforms that have been used widely in research and data collection and evaluation. Our findings will help to uncover data reliability problems and to propose changes in crowdsourcing platforms that can mitigate the inconsistencies of human contributions.

Keywords: crowdsourcing · task design· platforms.

1 Introduction

There are many successful examples on the web of crowdsourcing platforms. However, the features and services provided for the requesters vary from one platform to another, and no single platform meets all the possible requirements that the requesters may have.

We investigate the quality of the output of different platforms when the same task design and dataset is used. To study the reliability and consistency of the output of the platforms and to generalise the findings, we run a continuous evaluation of existing datasets and replicate the task over multiple weeks.

2 Related Work

Crowdsourcing platforms evaluation. In this context, a study by [3] attempts to validate Amazon Mechanical Turk (MTurk) as a tool for collecting data in cognitive behavioural research. They designed several types of experiments and compared the results with traditional laboratory ways of collecting data. The study showed that the quality of the data collected under the experimental conditions in MTurk is highly similar to the quality of the data collected the traditional laboratory way. A similar case study was presented by [1], who analysed the results of surveying the workers on their behaviour of using particular technologies. This research compared the results from MTurk and Survey Monkey to those obtained using a traditional survey. They demonstrated that crowdsourcing platforms can provide the same results and do it much faster when

compared to the traditional way of collecting survey data [1]. Despite some concerns related to the limitations of the technical and visual design of the task and unexpected behaviour such as dropping out of a task before finishing it, collecting data with crowdsourcing saves time and money and reach a wide range of users in a few seconds [3].

A few papers highlighted the differences between crowdsourcing platforms. In one of the recent studies, [6] introduced the new platform Prolific Academic (ProA) and compared the result of this platform with CrowdFlower (CF) and MTurk. The findings of this study recorded the highest response rate for participants in CF and the highest data quality for the participants in ProA and comparable to MTurk's [6]. Another study [5] used Rankings website to collect data and compare crowdsourcing platforms over two periods of time and according to a number of criteria: *type of service provided, quality and reliability, region, online imprint.* The findings of this study discuss the effect of the platforms characteristics of their traffic data and popularity [5, 4].

Time consistency of tasks. studies by [2] investigate the creation of evaluation campaigns for the semantic search task of keyword-based ad-hoc object retrieval using crowdsourcing task. They used a sample of entity-queries from the Yahoo! log and Microsoft log to evaluate the semantic search result. They prove that the reliability of crowdsourcing workers and the quality of the result was comparable to that of the experts even when repeating the same task over time [2]. Following this work, [8] extend the continuous evaluation of information retrieval (IR) systems using crowdsourced relevance judgments.

3 Research Questions

This research will address the following questions:

- **RQ1:** *Is there a significant difference in the quality, reliability, and consistency of the results for the same task repeated over a different time scale?*
- **RQ2:** *Is there a significant difference in the quality, reliability, and consistency of the results for the same task performed on different platforms?*

Answering RQ1 requires conducting a study where the same experiments will be repeated on a different time scale. We replicated the experiment using the same part of the dataset for the same assumption discussed in [2, 7] for measuring repeatable and reliable evaluation over crowdsourcing systems. These studies show experimental proofs that a crowdsourcing platform produces a scalable and reliable result over a repetition time of one month. We examined the consistency of the same task over a shorter time scale (once a week).

RQ2 offers an in-depth analysis and practical comparison of crowdsourcing platforms. We investigated the replication of the same task over multiple crowdsourcing platforms and over different levels of workers' experience and accuracy as provided by each platform. Two of the most popular platforms, that have

been used in crowdsourcing business and research studies of data evaluation and acquisitions, that is, Amazon Mechanical Turk (MTurk) and Figure Eight (F8), have been chosen for this study.

For both research questions and for each platform, we ran multiple types of tasks and measured the stability of the performance over the variations of the following factors:

- The quality of the task interface.
- The workers' experience level provided by the platform.

The evaluation of these factors depended on the completion time of the task and accuracy of the result. Moreover, with repeating the same task every week, the overall time of completing the batch on each platform will be recorded.

4 Experimental Results: Phase 1

The experiments in this phase used the plain interface similar to the one presented in [2]. We repeated the same experiment five times (once every week) and it was launched on the same day of the week and at the same time on each platform. Each task consisted of 20 tweets to be judged by 150 workers. The workers were rewarded with 0.15$ and they could do the task only once since after they finished they were excluded from participating in another batch of the task.

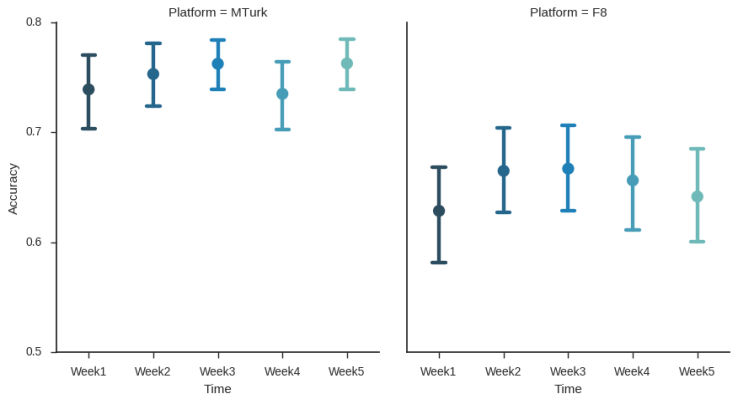

Fig. 1. Accuracy distribution over time.

Table 1 presents the results of the baseline phase experiments for the tweets dataset with comparison between the two selected platforms. The results show some consistency over the five runs on each platform. Workers were finishing the task faster in MTurk, where the average time per assignment was approximately 4 minutes, while it took approximately 6 minutes in F8. The overall accuracy

for each run on MTurk was more than 73% whereas it was in the range of 60% on F8 as shown in Figure 1. Although the results from MTurk are significantly better than those from F8, the total completion time for the whole batch took an average of 3 days in MTurk and 4 to 7 hours in F8.

Table 1. Results of five runs in MTurk and F8

	MTurk	F8
Average Time per Assignment	4 m, 16 s	6 m, 09 s
	4 m, 49 s	6 m, 33 s
	4 m, 24 s	6 m, 18 s
	4 m, 25 s	5 m, 30 s
	4 m, 37 s	5 m, 49 s
Avg.Accuracy & Standard deviation	0.73 ± 0.20	0.63 ± 0.28
	0.76 ± 017	0.66 ± 0.25
	0.76 ± 0.14	0.67 ± 0.25
	0.74 ± 0.19	0.66 ± 0.27
	0.76 ± 0.14	0.64 ± 0.28
Completion Time for the Batch	3 d, 00 h, 14 m	05 h, 11 m
	3 d, 01 h, 29 m	04 h, 45 m
	2 d, 08 h, 36 m	07 h, 10 m
	3 d, 13 h, 54 m	04 h, 43 m
	3 d, 03 h, 28 m	04 h, 04 m

A two-way ANOVA was conducted to examine the effect of repeating the same task several times and on two different platforms on the accuracy of the results (Table 2). There was a statistically significant interaction between the effects of repeating the task on different platforms on the accuracy $p < 0.05$. There were no differences between running the experiment several times on each platform which indicates the consistency of the outcome of each platform. We will investigate the reasons for having this difference accuracies.

Table 2. Results of 2 ways ANOVA test.

	sum_sq	df	F	PR(>)
C(Platform)	3.65	1.0	71.4	0.68e-16
C(Time)	0.17	4.0	0.84	0.49
C(Platform):C(Time)	0.08	4.0	0.42	0.80
Residual	76.17	1490.0	NaN	NaN

5 Future Directions

For the advanced phases of this study, we will investigate why we had these results in the first phase. One of the reasons could be the workers' diversity and

their level of experience. Another reason could be the variation of the amount of payment for different channels in F8. With this study, we hope to reach a reasonable level of understanding what are the best strategies and advise crowdsourcing users on the best way to achieve better service from the system.

6 Acknowledgements

This project is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 732328.

References

1. Bentley, F.R., Daskalova, N., White, B.: Comparing the Reliability of Amazon Mechanical Turk and Survey Monkey to Traditional Market Research Surveys. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17. pp. 1092–1099 (2017). https://doi.org/10.1145/3027063.3053335
2. Blanco, R., Halpin, H., Herzig, D.M., Mika, P., Pound, J., Thompson, H.S.: Repeatable and Reliable Search System Evaluation using Crowdsourcing. Journal of Web Semantics **21**, 923–932 (2011). https://doi.org/10.1016/j.websem.2013.05.005
3. Crump, M.J.C., McDonnell, J.V., Gureckis, T.M.: Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. PLoS ONE **8**(3) (2013). https://doi.org/10.1371/journal.pone.0057410
4. Mourelatos, E., Frarakis, N., Tzagarakis, M.: A Study on the Evolution of Crowdsourcing Websites. ISSNOnline) European Journal of Social Sciences Education and Research **11**(1), 2411–9563 (2017)
5. Mourelatos, E., Tzagarakis, M., Dimara, E.: A REVIEW OF ONLINE CROWDSOURCING PLATFORMS. South-Eastern Europe Journal of Economics **14**(1), 59–74 (2016)
6. Peer, E., Samat, S., Brandimarte, L., Acquisti, A.: Beyond the Turk : Alternative platforms for crowdsourcing behavioral research. Journal of Experimental Social Psychology **70**(January), 153–163 (2016). https://doi.org/10.1016/j.jesp.2017.01.006
7. Tonon, A., Demartini, G., Cudré-Mauroux, P.: Combining inverted indices and structured search for ad-hoc object retrieval. In: SIGIR. p. 125 (2012). https://doi.org/10.1145/2348283.2348304
8. Tonon, A., Demartini, G., Cudré-Mauroux, P.: Pooling-based continuous evaluation of information retrieval systems. Information Retrieval **18**(5), 445–472 (2015). https://doi.org/10.1007/s10791-015-9266-y