Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 9, August 2021: 1100 - 1109

Investigating Reliability and Stability of Crowdsourcing and Human Computational Outputs based on Artificial Intelligence

Zeeshan Rasheed [1],Naeem Ahmed Ibupoto[2]

ABSTRACT

Crowdsourcing and human computational outputs provide a scalable and convenient way to perform different human generated tasks use to evaluate the effectiveness of different crowdsourcing platforms based on artificial intelligence. The crowdsourcing generated datasets depend on multiple factors including quality, task reward and accuracy measuring filters. The present was designed to evaluate reliability and stability and consistency of crowdsourcing platforms outputs. We conduct a longitudinal experiment over specified time and two crowdsourcing platforms, Amazon Mechanical Turk and CrowdFlower to demonstrate how the outcomes of reliability varies considerably across platforms whereas repeating tasks over same platform yields consistent results. Different tasks on three different datasets were performed to evaluate the quality of the task interface, employees' experience level supplied by the platform and the evaluation of aquracy of outcomes dependent on the task completion time. The outcomes revealed significant ($p<0.05$) preeminence of MTurk over CrowdFlower in terms of reliability, accuracy and completion time taken for a task. The tasks replicated on these two platforms showed significant difference in quality based outcomes. The data quality of same repeated tasks over different time was stable in the same platform while it's was different for different crowdsourcing platform over differ time span. It was concluded from the findings that by employing standard platform crowdsourcing settings varying order and magnitude of task completion on different platforms can easily be achieved with varying levels of accuracy

Keywords: *Crowdsourcing, Human computation, Artificial Inteligence, MTurk, CrowdFlower*

Introduction

Human computation is a new and developing study topic that focuses on using human intellect to tackle issues that are beyond the reach of current Artificial Intelligence (AI) algorithms. Human computing systems may now use the talents of an unprecedented number of individuals over the Web to conduct complicated computations, thanks to the expansion of the Web. Human computation has resurfaced as an important study area as part of collective intelligence. The word "crowdsourcing" was developed shortly after, spawning yet another field of study closely related to human computation (Omar, 2019).

After over a decade of intensive study on human computation and crowdsourcing, numerous crowdsourcing-based techniques and commercial models for managing and distributing labour to the public have developed. In this respect, crowdsourcing (the most often used phrase) may be seen from

[1]Zeeshan Rasheed Mir Chakar Khan Rind University Sibi, zeeshanrasheed1992@yahoo.com,

[2]Naeem Ahmed Ibupoto Mir Chakar Khan Rind University Sibi, naeem.ahmed@mckru.edu.pk,

two angles: from a business domain-specific perspective and from a technological domain-independent one. In recent years, a substantial and thriving pay-based crowdsourcing business and workforce (crowdsortium.org) has arisen, which offers a wide range of activities over the Internet to a global population of on-demand, 24/7 employees (Djellel *et al.*, 2019). This involves annotating data to educate AI/ML systems and developing hybrid human-in-the-loop systems that use people to do computational tasks in real time. With collective intelligence, "wisdom of crowds," and crowd computing, groups of individuals and/or systems may solve issues that are beyond the capabilities of any single human or machine. Human-centered qualitative studies (Gang *et al.*, 2021).

The upsurge of many crowdsourcing platforms has made it possible to gather human labelling on a large scale. Researchers that use these platforms (called requesters) hope to get repeatable, reliable, and reproducible data from the crowd, as dictated by scientific best practise. We modify these conventional standards in scientific experiments in a crowdsourcing context. Reliable findings are produced when crowdsourced data is accurate when compared to standard data when other quality criteria, such as inter-annotator agreement, are used. One of the key problems in crowdsourcing is using quality control techniques to achieve accurate results (Mukasheva and Payevskaya, 2020).

Holding consistency after repeating the same experiment several times yields repeatable findings. When consistent observations can be made across multiple crowdsourcing platforms, reproducible findings may be produced. Previous research has looked at the external and internal variables that influence output variability among crowdsourcing platforms. Nonetheless, the possibility of replicating outcomes for similarl tasks across various platforms has never been investigated. Previous research in machine learning (Paul, 2017; Hasna El Alaoui El Abdallaoui *et al.*, 2020; Xiaohui *et al.*, 2020) has used reliability to express consistency. Only a few research have looked into the consistency of crowdsourcing results. As a result, several questions remain unanswered: 1) Does repeating the same job over the same dataset on the maching platform result in various levels of result quality? 2) Can the same job be released on multiple platforms (and therefore with possibly diverse audiences) and achieve the same degree of result quality?

There are several successful instances of crowdsourcing platforms on the internet. However, the features and services supplied to requesters differ from one platform to the next, and no one platform can fulfil all of the requesters' expectations. When the same task design and dataset are utilised, we examine the output quality of different platforms. We perform a continuous assessment of existing datasets and duplicate the job over many weeks to examine the dependability and consistency of the platforms' output and to generalise the findings (Benedikt *et al.*, 2019).

We provide the first experimental investigation in this work that shows how crowdsourcing outcomes are more or less compatible with scientific research criteria. We looked at how the identical job was replicated on two differnt platforms, MTurk (Amazon Mechanical Turk) and CrowdFlower (CF) to see how various degrees of expertise and accuracy were supplied by each platform.

Research Questions

Research Question 1 (RQ1): Is there a substantial variance in the reliability, quality and consistency of outcomes on performing recurring task over different time period?

Research Question 2 (RQ2): Is there a substantial variance in the reliability, quality and consistency of outcomes on performing the same task over different platforms?

METHODOLOGY

To address RQ1 and RQ2, we conducted different experiments. We conduct a longitudinal experimentation across two different crowdsourcing platforms, MTurk and CF to demonstrate how the result reliability varies significantly in different platforms whereas repetitive experiments over the same platform yields constant results.

To answer RQ1, you'll need to perform a research in which the identical trials are replicated on a diverse time frame. We repeated the experiment with the similar part of the task to test the same hypothesis for evaluating repeatable and reliable evaluation in crowdsourcing systems. These experiments demonstrate that a crowdsourcing platform may generate a scalable and trustworthy outcome after a month of repeating. We looked at how well the identical activity performed over a shorter period of time (once a week). RQ2 provides an in-depth look of crowdsourcing platforms as well as a practical comparison.

We looked at how the same task may be replicated across several crowdsourcing platforms, with varied degrees of worker experience and accuracy given by each platform. Amazon Mechanical Turk (MTurk) and CrowdFlower (CF), two of the most prominent platforms for crowdsourcing business and research studies of data assessment and acquisitions, were chosen for this study. We conducted numerous types of tasks for both the research topics and each platform, measuring performance stability across variations of the following factors: – The quality of the task interface. – The employees' experience level supplied by the platform. The evaluation of aquracy of outcomes were dependent on the task completion time.

Dataset

Three different types of labelling dataset used for the task design of the study were

1. Documents
2. tweets
3. images.

Task Design:

The task comprised of a batch of 10 documents as Dataset 1 and twenty documents from Datasets 2 and 3. The documents were obtained through sampling evenly at random from the datasets. Each week, three different Human Intelligent Tasks (HITs) were released on each platform. The quantity of papers was chosen to guarantee that each assignment could be completed in 5-6 minutes. The user interface has been created to seem the same on both systems. We hosted the task interface on an offsite server and utilized iframes to display it on each device. The main variation between the two systems' worker experiences was the way the task preview was displayed and how

workers could access the job. Both completion time and population selection bias may be affected by these variables. 3

Analysis of The Data

The obtained data was statistically analyzed by using two ways ANOVA at a 0.05 level of significance. Graph pad prism 9.1.2 software was used for the analysis.

RESULTS

To evaluate the accuracy and dependency of crowdsourcing platforms, different experiments were conducting using MTurk and CF platforms to evaluate the validity and accuracy of each task design for each plateform. Two sets of experiments were designed. In Experiment 1 same tasks were repeated for three datasets to evaluate reliability of two selected crowdsourcing platforms to cover RQ1. Experiment 2 was conducted to evaluate how the same task can replicated on two crowdsourcing platforms, with varied degrees of worker experience and accuracy given by each platform to address RQ2.

Experiment 1

Experiment 1 involved repeated experiment with the same part of the dataset to test the same hypothesis for evaluating repeatable and reliable evaluation in crowdsourcing systems. For this purpose the same experiment was repeated five times (each task once/week) for each platform involving three different data sets. Three major tasks for three datasets were evaluated to compare these two crowdsourcing platforms.

1. Average time for an assignment
2. Average Accuracy
3. Time of completion per Batch

Average time for an assignment

Average time taken by two selected platforms to perform an assignment for three showed significant difference ($p<0.05$) between two platforms (Table 1). The outcomes revealed consistency of both the platforms for the five runs however MTurk found to be more faster than CF. It was found that MTurk workers procured average 4 minutes to complete dataset1 task, average 4.6 minutes to complete task of dataset and average 6 minutes to complete dataset task 3 (Table 1). CF workers took average 7, 6.6 and 7.2 minutes to complete repetitive tasks of dataset1, dataset 2 and dataset 3 respectively. Overall faster task performing activity was observed for MTurk as compared to CF for all three data sets.

Table 1. Average Time Taken by MTurk and CF for Repeating Runs of Three Datasets

Data Sets	Interval	MTurk	CF	F	p-Value
Data set 1	Week 1	4 m, 13 sec	7 m, 12 s	1068	0.0003*
	Week 2	4 m, 45 sec	7 m, 36 s		
	Week 3	4 m, 20	7 m, 23 s		

	Week 4	4 m, 27	6 m, 23 s		
	Week 5	4 m, 40	6m, 52 s		
Data set 2	Week 1	5 m, 18 sec	6 m, 9 s	434.6	0.0001*
	Week 2	5 m, 53 sec	6m, 16 s		
	Week 3	5 m, 30	7 m, 20 s		
	Week 4	4 m, 32	6 m, 25 s		
	Week 5	4 m, 37	6m, 34 s		
Data set 3	Week 1	6 m, 23 s	7 m, 29 s	152.7	0.0001*
	Week 2	6 m, 35s	7m, 34 s		
	Week 3	6 m, 34 s	7 m, 25 s		
	Week 4	6 m, 12 s	6 m, 22 s		
	Week 5	6 m, 13	6m, 13 s		

*Significant value $p<0.05$

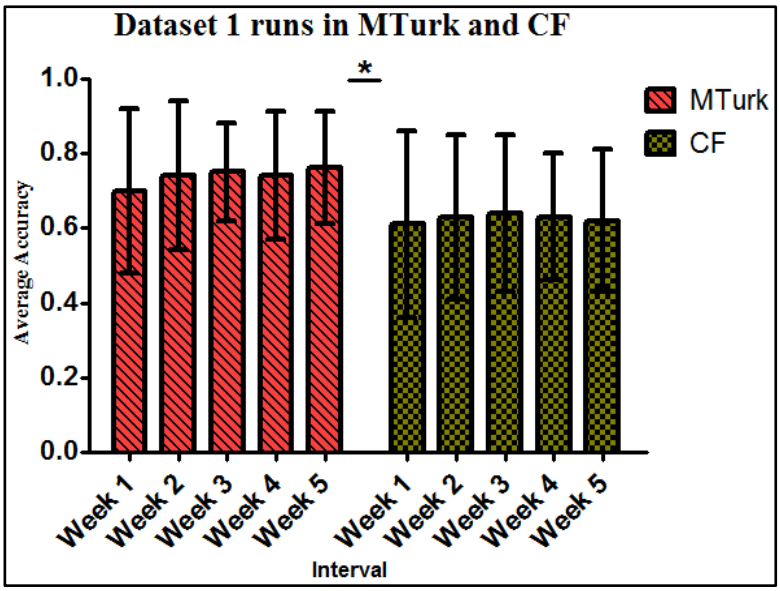
Average Accuracy

The outcomes of baseline experiments conducted for average accuracy of two platforms for three different datasets are presented in Table 2. High level of quality consistency was observed for both the platforms over five repeating tasks for three datasets however the percentage of accuracy for MTurk was significantly high as compare to CF ($p<0.05$). Overall 74% average accuracy for dataset1 (Table 1, Figure 1), 68% for dataset 2 (Table 1, Figure 2) and 67% for dataset 3 (Table 1, Figure 3) was observed on MTurk while for CF the average accuracy observed were 63%, 61.2% and 61% for data set 1, dataset 2 and dataset 3 respectively.

Table 2. Average Accuracy of MTurk and CF for Repeating Runs of Three Datasets

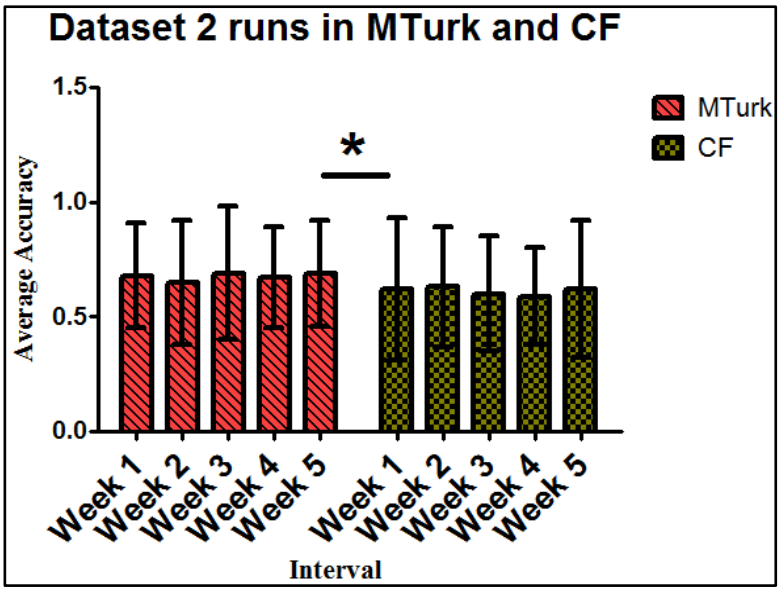
Data Sets	Interval	MTurk	CF	F	p-Value
Data set 1	Week 1	0.70 ± 0.22	0.61 ± 0.25	161.7	0.0001*
	Week 2	0.74 ± 0.20	0.63 ± 0.22		
	Week 3	0.75 ± 0.13	0.64 ± 0.21		
	Week 4	0.74 ± 0.17	0.63 ± 0.17		
	Week 5	0.76 ±0.15	0.62 ±0.19		
Data set 2	Week 1	0.68 ± 0.23	0.62 ± 0.31	344.9	0.0001*
	Week 2	0.65 ± 0.27	0.63 ± 0.26		
	Week 3	0.69 ± 0.29	0.60 ± 0.25		
	Week 4	0.67 ± 0.22	0.59 ± 0.21		
	Week 5	0.69 ± 0.23	0.62 ±0.30		
Data set 3	Week 1	0.68 ± 0.31	0.62 ± 0.21	166.2	0.0001*
	Week 2	0.65 ± 0.29	0.63 ± 0.28		
	Week 3	0.69 ± 0.26	0.60 ± 0.28		
	Week 4	0.67 ± 0.32	0.59 ± 0.31		
	Week 5	0.69 ±0.31	0.62 ±0.22		

*Significant value $p<0.05$



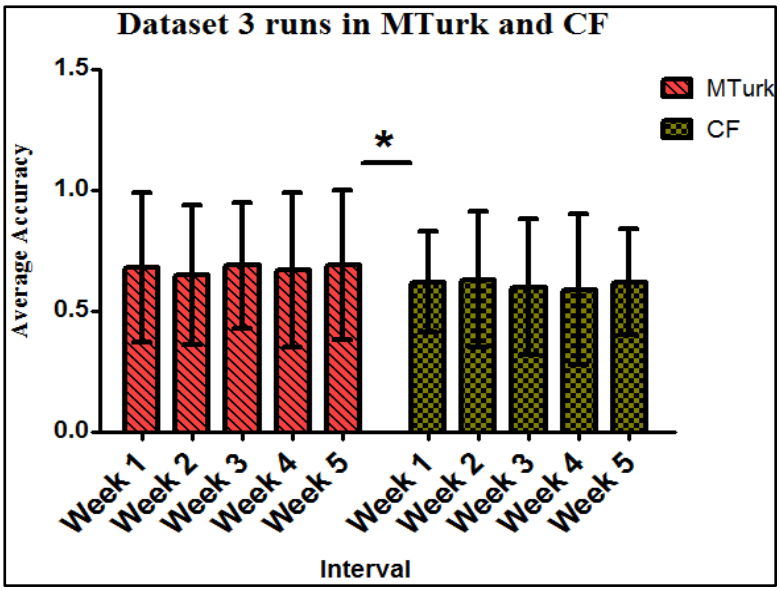
*** Significant difference**

Figure 1: Average accuracy of MTurk and CF platforms for dataset 1.The figure depicts significant high average accuracy of reparative runs during 5 weeks on MTurk as compare to CF.



*** Significant difference**

Figure 2: Average accuracy of MTurk and CF platforms for dataset 2.The figure depicts significant difference in average accuracy of reparative runs during 5 weeks on MTurk as compare to CF.



*** Significant difference**

Figure 3: Average accuracy of MTurk and CF platforms for dataset 3.The figure depicts significant difference in average accuracy of reparative runs during 5 weeks on MTurk as compare to CF.

Time of Completion per Batch

The average completion time for the complete batch observed for MTurk was 3 days 7 hrs for dataset 1, 22 hrs for dataset 2 and 7 days 8 hrs for dataset 3. For CF crowdsource platform time of completion per batch was 6 hrs, 12 hrs and 2 days 19 hrs for dataset 1, dataset 2 and data set 3 respectively. The outcomes revealed that completion of repeating tasks was significantly ($p<0.05$) faster on CF as compare to the MTurk (Table 3).

Table 3. Time of completion per Batch on MTurk and CF for Repeating Runs of Three Datasets

Data Sets	Interval	MTurk	CF	F	p-Value
Data set 1	Week 1	3d, 01h, 29m	07h, 13m	991.2	0.0001*
	Week 2	3d, 02h, 31m	07h, 22m		
	Week 3	3d, 09h,12m	04h, 39m		
	Week 4	3d, 12h,55m	05h, 21m		
	Week 5	3d, 08h, 13m	05h, 11m		
Data set 2	Week 1	18h, 21m	10h, 31m	65.22	0.0002*
	Week 2	23h, 11m	12h, 12m		
	Week 3	22h,28m	09h, 32m		
	Week 4	24h,15m	14h, 16m		
	Week 5	24h, 43m	14h, 19m		
Data set 3	Week 1	6d,06h, 23m	2d, 10h, 11m	123.7	0.0001*
	Week 2	7d,5h, 21m	2d, 23h, 22m		

	Week 3	6d, 2h,20m	2d, 19h, 31m		
	Week 4	7d,12h,25m	1d, 22h, 36m		
	Week 5	7d,14h, 42m	2d, 18h, 39m		

*Significant value p<0.05

Experiment 2

ANOVA (two way) was performed to evaluate the outcomes of repetitive performance of the same task on two varying platforms several times. Significant interaction was observed between accuracy and repeated tasks on two platforms as $p < 0.05$. however the outcomes also revealed that there is no effect of consistency of each platform on the experiment running several time per each platform (Table 4).

Table 4: Two way ANOVA of repetitive same tasks on MTurk and CF for three datasets

Variables	Sum_Seq	df	F	p-Value
Platform	0.12	1.0	1.7	0.02*
Week	0.09	1.0	4.2	
Accuracy **Platform-week**	0.002	1.0	0.1	
Platform	0.05	1.0	2.31	0.08
Week	.14	1.0	1.21	
Consistency **Platform-week**	0.15	1.0	3.21	

*Significant value p<0.05

DISCUSSION

The present study covers the outcomes under two research questions in two phases. In phase one (Experiment 1) repeated experiment with the same part of the dataset to test the same hypothesis for evaluating repeatable and reliable evaluation in crowdsourcing systems in terms of average time taken for a task, accuracy, and completion time . Experiment 2 in phase 2 was conducted to evaluate how the same task can replicated on two crowdsourcing platforms, with varied degrees of worker experience and accuracy.

The outcomes of experiment 1 revealed persistent preeminence of MTurk over CF in terms of reliability, accuracy and completion time taken for a task. One possible elucidation of these outcomes is basic quality control measure provided by both MTurk and CF crowdsourcing platforms (Brian *et al.*, 2021). Workers get paid through CF crowdsource even if the quality of the task is not pleasing while the workers have options to rejection payment for a task performance in MTurk (Daniel *et al.*, 2019). Moreover there is no quality control measures implanted in CF system (Good *et al.*, 2015) the workers can have easy access to all information in this paltform as compare to the MTurk. In this way low quality control measures in CF crowdsource system reduces workers performance, reliability and accuracy of this crowdsourcing platform.

The significant difference in average time/ assignment was also observed between both platforms which could be related to the demographic and language distribution of the workers crowd

of the platforms. The majority Of MTurk users are from USA using native English language and are more conversant with the data and tasks items presented as tweets leading to faster task completion as compare to CF workers who are diverse group demographically (Irene *et al.*, 2019).

The results of experiment two involving repetitive performance of the same task on two varying platforms several times revealed significant effect of repeating tasks on the accuracy on two platforms as $p < 0.05$ which was in line with the work of () demonstrating similar results. the outcomes also revealed that there is no effect of consistency of each platform on the experiment running several time per each platform which was contrary to the findings of (Sujoy and Malay, 2017; Maria *et al.*, 2018; Antoine *et al.*, 2020). Insignificant relation of platform versus repetitive task performance could only be due to the fact that we mainly run three datasets based on simple classification. We did not tests these crowdsourcing platforms for combined set of complex datasets in our study.

CONCLUSION

This study evaluated the comparative efficacy of two crowdsourcing platforms MTurk and CF by repeating and reproducing different tasks. It was concluded from the findings that by employing standard platform crowdsourcing settings varying order and magnitude of task completion on different platforms can easily be achieved with varying levels of accuracy and by rescaling quality control measures both reproducibility and repeatability can equally be achieved by different crowdsourcing platforms.

. REFERENCES

1. Antoine T, Damien D, Julian A. 2020. he Colectyng Model for the Evaluation of Game-Based Learning Activities. Games and Learning Alliance. 32:401-407.
2. Benedikt M, Juho H, Alexander M. 2019. Cooperation or competition – When do people contribute more? A field experiment on gamification of crowdsourcing. International Journal of Human-Computer Studies 127:7-24.
3. Brian D, Ondov FY, Matthew K, E. N, F. S. 2021. Revealing Perceptual Proxies with Adversarial Examples. IEEE Transactions on Visualization and Computer Graphics 27:2:1073-1083.
4. Daniel S, Dubravka CK, Benjamin H. 2019. Ethical norms and issues in crowdsourcing practices: A Habermasian analysis. Information Systems Journal 24:4:811-837.
5. Djellel D, Alessandro C, Gianluca D, Cudré-Mauroux. 2019. Deadline-Aware Fair Scheduling for Multi-Tenant Crowd-Powered Systems. ACM Transactions on Social Computing 2:1:1:29.
6. Gang W, Zhiyong C, Jia L, Donghong H, Baiyou Q. 2021. Task assignment for social-oriented crowdsourcing. Frontiers of Computer Science. 15:2:218-223.
7. Good BM, Nanis M, Wu C, Su AI. 2015. Microtask crowdsourcing for disease mention annotation in PubMed abstracts. Pac Symp Biocomput.282-293.
8. Hasna El Alaoui El Abdallaoui, Abdelaziz El Fazziki, Mohamed S. 2020. Crowdsourcing and Blockchain-Based E-Government Applications: Corruption Mapping. Smart Applications and Data Analysis. 76:86-99.
9. Irene C, Gloria Re C, Andrea F. 2019. Refining Linked Data with Games with a Purpose. Data Intelligence. 8:1-26.

10. Maria AB, Irene C, Andrea F, Monia EM, Vijaycharan V. 2018. A crowdsourcing-based game for land cover validation. Applied Geomatics. 10:1:1:11.
11. Mukasheva MU, Payevskaya YV. 2020. Semantic influence of programming on the development of thinking of students: background, research and prospects. Open Education. 24:1:45.
12. Omar A. 2019. The Practice of Crowdsourcing. Synthesis Lectures on Information Concepts, Retrieval, and Services 11:1:1:149.
13. Paul S. 2017. Situating Machine Intelligence Within the Cognitive Ecology of the Internet. Minds and Machines. 27:2:357-380.
14. Sujoy C, Malay B. 2017. Judgment analysis of crowdsourced opinions using biclustering. Information Sciences. 375:138-154.
15. Xiaohui W, Dion Hoe-Lian G, Ee-Peng L. 2020. Understanding Continuance Intention toward Crowdsourcing Games: A Longitudinal Investigation. International Journal of Human–Computer Interaction. 36:12:1168-1177.