

# Report on implemented crowdsourcing task in Amazon Mechanical Turk

## Task specification

The task requirement was to ask crowd workers to label the relatedness of two pieces of text. According to this specification, I decided to present workers with two pieces of text and ask them to choose whether the first text describes the second one. The worker can choose only from two options: “Yes” or “No”. The idea of the task is to check whether the human emotion expressed in the second text is also captured in the first one. An example of such a task can be found below:

Does the sentence in Text 1 correctly describe the sentence(s) in Text 2?

**Text 1:** The person is experiencing or expressing happiness

**Text 2:** I love the sunshine and the happiness it brings

1

(Number key for revealing the answer to the voluntary question at the bottom of the page. You have to collect the keys from all 20 tasks in order to unlock the answer to the question. Please view the full instructions for further detail.)

Select an option

Yes	1
No	2

In this specific case the worker should select option “Yes”, as Text 1 correctly describes Text 2 (this example is trivial, as the word “happiness” is contained in both texts).

## Used data

The sentences contained in Text 2 come from a data set gathered on Crowdfunder<sup>1</sup>, where workers were asked to label thousands of tweets for their emotional content. However, the quality of the assigned labels is not very high and there are many tweets that have obviously been wrongly classified. Therefore, in order to improve the quality of the data, a selection of the labelled tweets can be processed again through a crowdsourcing platform. Instead of asking users to choose one of 13 different categories for the tweets, as in the Crowdfunder assignment, this time one can just ask workers to check whether the Crowdfunder label has been correctly assigned to the tweet. The Crowdfunder label would be contained in Text 1, whereas the tweet itself would be contained in Text 2. I have chosen Amazon Mechanical Turk to implement the new task.

## Task design

### ▪ General information

I have decided to gather 20 different tasks and post them all in a single batch. This way workers can get compensated for the initial time overhead they invest to understand what is required from them and get used to the task design. In addition, workers should be more willing to go through all tasks due to the curiosity-provoking measure I have implemented (explained below).

---

<sup>1</sup> <https://data.world/crowdfunder/sentiment-analysis-in-text>

The award for the correct submission of a task has been set to 0.05 USD. Assuming that the completion of the whole batch would take approximately 10 minutes on average, this award should be competitive to other tasks on Amazon Mechanical Turk.

The number of workers that have to process the tasks has been set to only 9, in order to avoid the additional Amazon fee surcharge<sup>2</sup>. However, this amount could always be increased, depending on the budget available. Responses from more workers would of course lead to more robust aggregated results.

Not every worker is allowed to work on the tasks. As the tweets contain many English colloquial phrases, the tasks have been limited to only US-based workers. Workers on Amazon Mechanical Turk come predominantly from the US and India (Difallah, Filatova, and Ipeirotis, 2018). Therefore, it makes sense to only concentrate on US workers. In order to benefit from the peak in US-based worker activity, the tasks should be posted at midnight UTC time, following the findings of Difallah et al. Moreover, only trustworthy workers with more than 50 successfully submitted tasks and a task approval rate of more than 80% will be considered.

- Quality control mechanisms

Two “gold standard” tasks have been included in the batch of 20 tasks. An example of such a “gold standard” task is shown in the screenshot on Page 1 of this report. The emotion in Text 1 has been explicitly mentioned in Text 2 and therefore it is obvious that diligent workers should choose option “Yes”.

Through those two questions we can filter out responses from workers, who are “fast deceivers” and possibly even “smart deceivers” (Gadiraju, Demartini, Kawase, and Dietze, 2019). The feedback of workers who gave a wrong answer to both questions will not be considered in the final aggregation of results. Those workers should also not receive any reward, as their feedback is very likely just spam.

- User engagement technique

Following the idea of Law, Yin, Goh, Chen, Terry, and Gajos (2016), a curiosity-provoking measure has been introduced in the task design. In order to motivate workers to complete more tasks, they are asked to guess from which popular social platform the sentences contained in Text 2 come from. People are made aware of a salient information gap in their knowledge, as they have to make a guess. Furthermore, workers are provided with means to close the gap, as they process more Text 2 sentences and have the opportunity to find out what the correct answer is. As shown in the screenshot on Page 1, each task in the batch contains a number key. By collecting these number keys, workers can unlock an Extra Task, which will reveal the name of the social platform. Workers would have to process all 20 tasks, if they want to be sure that they successfully unlock the correct answer. In theory, workers could just preview the 20 tasks, without actually working on them. This is however of no benefit for the workers, as in that case they do not get rewarded and simply clicking through all tasks is relatively time-consuming.

---

<sup>2</sup> Tasks with 10 or more assignments will be charged an additional 20% fee on the reward paid to workers: <https://requester.mturk.com/pricing>

## Literature references

Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and Dynamics of Mechanical Turk Workers. In Proceedings of the Eleventh ACM International Conference of Web Search and Data Mining (WSDM '18). ACM, New York, NY, USA, 135-143.

Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2019. Crowd Anatomy Beyond the Good and Bad: Behavioral Traces for Crowd Worker Modeling and Pre-selection. *Computer Supported Cooperative Work* 28(5): 815-841.

Edith Law, Ming Yin, Joslin Goh, Kevin Chen, Michael A. Terry, and Krzysztof Z. Gajos. 2016. Curiosity Killed the Cat, but Makes Crowdtwork Better. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16). ACM, New York, NY, USA, 4098-4110.