Crowdsourcing and Text Highlighting to Improve Text Analysis

Abstract

Crowdsourcing is widely used to solve problems difficult for computers [17]. The most common tasks on crowdsourcing platforms are information finding and verification, content interpretation and creation, sentiment analysis, and surveys [18, 19, 20]. **In our study, we aim to verify if the crowd can solve challenging problems of general text analysis**. We hypothesize that crowdsourcing applied to text analysis yields better results in comparison to the approach when one expert performs the entire task. We created three tasks of different complexity. Each task contains a large text and a binary question related to its content.

Related work

Borromeo et al. [12] studied how the task complexity influenced the quality of crowdsourced work. They observed that the accuracy of difficult tasks significantly decreased. In general, to overcome the complexity, difficult tasks are decomposed into independent micro-tasks. Those micro tasks can be performed sequentially or in parallel. For example, in Soylent [13], a text was divided into small pieces, and each of them was improved by several crowd workers. However, this kind of text decomposition is not always feasible for tasks involving text analysis because coherent information can be scattered across different paragraphs. Therefore, we designed a workflow where a task is performed by two workers in sequence. The first worker reads the text and highlights information relevant to the question. The second worker observes the highlights and provides the final answer. In our research, we study if the designed workflow can improve the accuracy of difficult crowdsourced tasks.

In crowdsourcing, an iterative workflow is popular. It was successfully applied in a number of applications such as creative writing, design, brainstorming, image transcription etc. [14, 15, 16]. Little et al. [16] compared parallel and iterative workflows and showed that the letter process with a sufficient number of iterations produced better results. However, with two iterations they achieved only a slight improvement. **As our workflow involves two participants and is similar to the iterative process with two iterations, it is interesting to check if we can gain better accuracy for text analysis tasks**.

In a sequential process of task execution, when the results of one worker are passed to the next one, the question arises if the entire text should be provided to the second worker or only its highlighted portions. The support for the former approach can be found in [7, 9, 10]. Farzan and Brusilovsky [7] demonstrated that **interpersonal trust affects the perception of the usefulness of annotated text**. Thus, providing the full text to the second worker impacts his or her trust and decision about the work of the first participant.

Another important question is how to transfer the results of work from one participant to another. In our research, we focus on highlights. Both workers are provided with a long and complicated text. The first worker is asked to mark all information in the text which can help the second worker to deduce the answer. The second worker is asked to respond to a question and find the answer in a highlighted text. The first participant highlights the text during the reading and thus employs the active text-marking strategy. The second participant uses passive highlights.

The influence of typographical cues on a reader has been extensively studied since the 1970s. Lorch in [3] suggests that the highlights influence on a reader's comprehension, memory, and information

search. Our tasks involve all these processes. **However, the existing research on the effectiveness of highlights on those processes has produced mixed results**. Some studies support the 1 idea that highlights improve comprehension, memory, and search; and others show the negative effect. It relates to both active and passive highlights. Moreover, there is no a substantial amount of research in the digital reading context.

A brief overview of the existing studies of the effects of highlights on comprehension and memory can be found in [4]. The existing research primarily estimates that effect by testing participants after they have read a text. However, in our settings, the workers knew in advance the question which they should answer using the text. Thus, in our tasks, the search process plays an important role. We rely on the von Restorff isolation effect and expect that the highlighted text attracts the readers' attention, thus impacting their memory, comprehension, and search behavior. Chi and Hong [6] in their eye-tracking study proved that a reader is more likely to pay attention to the highlighted sentences. In their experiments, the participants were presented with several questions and were asked to find the answers in a text. Users, who were presented with the highlighted text, demonstrated better accuracy and response time. On the contrary, Farzan and Brusilovsky [7, 11] reported that the participants in their study were only slightly influenced by highlights while performing a text search. Although, the provided passive highlights were not entirely related to the task.

Given a controversial overall effect of highlighted text on a reader, a number of studies further explore its influence on different types of subjects. **They indicate the substantial variation based on the participants' reading abilities, cognitive styles, or age [4, 5]. The most definite conclusion relates to the detrimental impact on a reader of poor and irrelevant highlights [4, 5, 8]**. Additionally, Lorch [3] advocates the proportion of highlighted text should not be too high. Following this hypothesis, Dodson et al. [5] used 15% in their tasks. As it relates to our settings, the above findings suggest the high importance of the results produced by the first worker who annotates the text. Given mixed results of the effect of highlights, we aim to establish whether it is possible to use highlighted text to effectively transfer the results of work from the first user to the second one. Additionally, the described above studies used laboratory settings, while we recruit workers on a crowdsourcing marketplace. We hypothesize that it is possible to utilize a crowdsourcing approach in order to perform complex text analysis. Furthermore, we designed a sequential task workflow when workers perform different assignments. We check if it is possible to improve task performance in terms of accuracy and time spent by workers without redundant assignments.

Summarizing the listed above findings we identify a set of challenges that we face in our study:

- The crowdsource workers who work on the same task don't know one another. Thus the second worker may not trust the work that was performed by the first one. To mitigate this we provide the second worker with the full text. Thus they could see the highlights in context.
- Our tasks are completed within two steps. However, two iterations may not suffuse. To account for this we create relatively simple tasks that can be performed within two iterations.
- The highlights' effectiveness is based on the participants' cognitive styles and other personal characteristics. To account for this, we ask workers to fill in a short demographic survey after completing the task. Additionally, before letting workers start the task, we screen them with a preliminary test and filter out those who failed it.

Keeping this in mind **we aim to verify the hypothesis that crowdsourcing applied to text analysis yields better results in comparison to the approach when one expert performs the entire task. We use text highlights as a medium for passing work between the task stages.**

Tasks description

We designed three tasks. Each task consists of a text and a question regarding its content. The tasks involve a worker's comprehension, memory, and information search processes. For each task we

compiled three different texts. Each text is a 2-page description of a ski-resort: Andorra, Kitzbuel, or Bad Gastein. These are the real ski-resorts located in Europe, thus most probably unfamiliar for the US Amazon Mechanical Turk workers. For each ski-resort we combined the information from its official web site with a general description from the web sites for travelers . By doing so, we interleaved irrelevant descriptions with data pertinent to the task. We created our texts in such a way that each of them contained the following information: (1) famous people who have a relation to a ski-resort; (2) different spots which a tourist can visit; (3) prices for tickets, hotels, meals, ski-passes etc.

We range our tasks by tree levels of difficulty. The most difficult task asks a participant to calculate a budget which is required to go on vacation. The easiest task asks to count the famous people mentioned in the text. The third task, which difficulty lies in between, asks if events mentioned in the text are mutually exclusive.

We hypothesize that splitting the tasks between two workers improves the accuracy. In our experiments, we assigned tasks to three groups of workers. The participants of the control group performed the entire task without the ability to highlight the text. In Treatment 1 group one participant performed the whole task but highlighting was allowed. In Treatment 2 group, each task was assigned to two workers: the first worker (annotator) highlighted the text and the second workers (labeler) observed the highlights and provided the final answer.

Initially, each task contained one question with a binary answer (YES or NO). However, this type of answer was considered as not sufficiently resistant to the dishonest workers who try to cheat the task. So, we assumed that the obtained results could be very noisy. Thus, we modified our tasks and asked workers to provide numerical answers. **In this paper, we compare the obtained results for binary and numerical tasks**.

To filter out the results of unscrupulous workers who did not put an effort into the task and possibly guessed the answer we applied a time threshold. For a task performed by a single worker we filtered out the answers which were given for less than a minute. For a task which was split between two workers the time threshold was 30 seconds for each worker. We applied different time thresholds but the difference in accuracy remained approximately the same. Thus, for all our tasks we used the above threshold values.

For binary answers, we filtered out 8% of the results for the Famous task, 16% for the Events task, and 23% for the Budget task. This percentage agrees with the task difficulty which we intended to impose. These percentage may see low, in comparison with data reported by Vuurens et al. In [25] they classified only 55% of all workers on MTurk as the proper ones. However, **in our study, before participants could proceed to a task, they should complete a quiz and a preliminary task on a test paragraph**. Those actions helped us to filter out some portion of spammers. We examined the number of filtered data for each task, text, treatment, and answer type (YES or NO). However, we did not observe any relation between those parameters and the number of filtered results. For the tasks with numerical answers, the overall number of discarded results is less than 20.

Worker incentives

In our experiments we implemented the following approach to incentivize workers. All workers in Control and Treatment 1, and labelers in Treatment 2 groups got bonuses if their answers were correct. Annotators from Treatment 2 group, i.e workers who only highlighted the text got bonuses if their corresponding labelers gave correct answers. All workers were informed about payments and bonuses before starting the task.

Tasks analysis

We begin by providing the overall results. Table 1 contains accuracy of the three tasks with binary answers. Table 2 shows the accuracy of numerical answers if we consider them as being correct only if they are below or above the corresponding thresholds of the binary questions. Table 2 omits the results for the Events task because we cannot directly map it to Table 1. In our binary questions we ask workers if it is possible to visit all the events, but in the numerical questions, we ask to provide the number of those events. Table 3 contains mean absolute relative errors, that is workers' errors relative to true answers.

Text (difficulty)	**Control correct/total (%)**	**Treatment 1 correct/total (%)**	**Treatment 2 correct/total (%)**
Famous (easy)	99/144 (69%)	102/158 (65%)	95/142 (67%)
Events (normal)	23/39 (59%)	20/33 (61%)	35/48 (73%)
Budget (difficult)	19/34 (56%)	34/44 (77%)	34/45 (76%)

Table 1: Accuracy of the tasks with binary answers.

Text (difficulty)	**Control, %**	**Treatment 1, %**	**Treatment 2, %**
Famous (easy)	67%	72%	71%
Budget (difficult)	82%	77%	74%

Table 2: Accuracy of the tasks with numerical answers.

Text (difficulty)	**Control**	**Treatment 1**	**Treatment 2**
Famous (easy)	0.189	0.208	0.231
Events (normal)	0.358	0.323	0.255
Budget (difficult)	0.289	0.268	0.290

Table 3: Mean absolute relative error.

Tables 5 and 6 represent median time spent by workers in our tasks.

Group of workers	**Control, mins**	**Treatment 1, mins**	**Treatment 2 (annotators), mins**	**Treatment 2 (labelers), mins**
Famous	3.6	4.2	4.9	2.8
Events	2.7	3.8	4.4	2.5
Budget	3.7	4.9	5.7	3.3

Table 5: Median time for all binary tasks.

Group of workers	**Control, mins**	**Treatment 1, mins**	**Treatment 2 (annotators), mins**	**Treatment 2 (labelers), mins**

Group of workers	Control, mins	Treatment 1, mins	Treatment 2 (annotators), mins	Treatment 2 (labelers), mins
Famous	4.2	5.9	4.7	2.8
Events	4.7	6.2	5.8	3.3
Budget	7	9.2	6.7	5.3

Table 6: Median time for all numerical tasks.

One main observation is that the **labor division reduces the time spent by the last worker on the task**. For all tasks the total time of both workers in Treatment 2 group is higher than of a single worker in Treatment 1. However, on average, a labeler spent significantly less time than a single worker who completed the entire task. In medical settings, a labeler can be viewed as a highly skilled nears who analyses a medial chart. Tables 5 and 6 suggest that the time which she spends for a chart analysis can be reduced if the text is preprocessed and the relevant information is highlighted. Interestingly, the median time for the binary Famous task is higher than for the Events task, disregarding of our intention to make the latter more difficult. On the other hand, it does not hold for the numerical tasks. Therefore, this difference can be attributed to higher number of guesses made by workers for more challenging tasks.

The effect of highlights is somewhat mixed, so we investigate it further in the following subsections. Next, we provide the analyses of each task: Famous, Events, and Budget. For each task, we present the results for each text: Andorra, Gastein, and Kitzbuel. Each text contains “clues” and “traps”. A clue is part of the correct answer. A trap is an unrelated object mentioned in the text and aims to mislead a careless worker.

Famous Task

We begin by the simplest Famous task which asks to count in the text only those famous people who have some relation to the ski resort. In the binary variant we ask it the number of those people is strictly less than 16. In the numerical version the workers need to submit the exact number of those persons. Table 7 describes each text.

Text	Number of clues	Number of traps	Correct answer
Gastein (easy)	15	0	Yes
Kitz (normal)	16	3	No
Andorra (difficult)	13	6	Yes

Table 7: Text description

The Gastein text is the easiest one. It does not contain irrelevant or repeated names. In the Kitzbuel text we repeated a few names and also mentioned a person who has no relation to the resort. However, even if a worker counted those false names it should not change the final answer because the number of correct names already exceeds 16. The Andorra text contains ten false names. We believe that five of those are easy to identify: they are general people who won a ski pass. Another two names belong to artists who composed songs about Andorra. The remaining false names belong to the prominent Andorran athletes and a radio magnate but they should be excluded from consideration because those people were already mentioned earlier in the same paragraph. The task states that each perrson should be counted only once, and we check in the quiz if a participant understands that. Tables 8 contains the accuracy of the binary Famous task, and Table 9 provides mean absolute relative

error for numerical answers. Recall that in Treatment 1 group one person highlighted the text and answered the question, and in Treatment 2 we used highlighting with labor division.

Text	**Correct Answer**	**Control correct/total (%)**	**Treatment 1 correct/total (%)**	**Treatment 2 correct/total (%)**
Gastein	Yes	46/46 (100%)	46/52 (88%)	47/49 (96%)
Kitzbuel	No	34/49 (69%)	39/52 (75%)	35/48 (73%)
Andorra	Yes	19/49 (39%)	17/54 (31%)	13/45 (29%)

Table 8: Accuracy of each text for the binary Famous task

Text	**Correct Answer**	**Control**	**Treatment 1**	**Treatment 2**
Gastein	15 names	0.137	0.146	0.061
Kitzbuel	16 names	0.129	0.188	0.120
Andorra	13 names	0.299	0.370	0.321

Table 9: Mean absolute relative error of each text for the numerical Famous task.

We observe that the workers' performance agrees with the designed difficulty of the texts. The accuracy of the Andorra text is the lowest. The other two tasks are easy and contain much more correct answers. The analysis of the results for each text is provided below:

1. In case of Gastein, Table 7 suggests that if all names were highlighted then a worker should provide the correct answer. We manually checked all incorrect answers, and in all cases the highlighting was correct (15 names were highlighted); however the participants reported that that number was bigger than 16. One reason of such an error could be that part of a highlighted name was moved to a new row and looked like two names. We can say that all three groups performed equally well. Gastein task was easy enough to perform it without highlighting or labor division.
2. Kitzbuel. For this text, the more names were counted the better, even if workers counted traps. The accuracy of Treatment 1 and Treatment 2 groups looks similar. It looks like highlighting helped workers to identify names in the text. However, workers did not analyze what names they highlighted. In particular:

- In Treatment 1 group, 94% of highlights contained traps. And those workers did not stop to highlight as far as they reached 16 names: 71% of them highlighted more than 16 names. There was only 1 worker who identified all clues and did not highlight traps.
- Treatment 2 highlights are more accurate: 87% of workers highlighted traps, and 64% of those highlighted more than 16 names; 4 workers identified all clues and did not highlight traps. Given a better accuracy of highlights in Treatment 2 group **we advocate that division of labor helps to increase the quality of work for participants who perform intermediate tasks**. It may be caused by additional bonuses to annotators for correct final answers given by labelers, or by increased personal responsibility. We checked if the labelers followed highlighters: 73% of labelers provided the answers consistent with highlights. Additionally, Table 5 and 6 show the reduced time that labelers spent on their tasks. Both findings suggest that **labor division helps with tasks asking to count relevant objects in the text. However, labelers performed some text analysis besides relying on the annotators' highlights** because even highlighted traps didn't impede labelers from providing a correct numerical answer. Moreover, 25% of the texts contained

excessive highlighting, but in those cases labelers performed equally well. In this report when we refer to the excessive highlighting we mean that more than 15% of the text was highlighted.

3. Andorra. To answer this task, workers had to correctly identify the majority of clues and traps. In both treatments the great majority of workers highlighted irrelevant names, for example, the lottery winners or Saint Nicholas. It may be that the workers who answered correctly scanned the text twice: first, they highlighted the names; next, they counted only the relevant persons. The control group performed the best, however, we do not know what tools they utilized to track names in the text. It may be that the workers in control group used a text file to gather and count unique names. The results suggest that the highlights were not the best tool to successfully perform on the complex text about Andora. For Andorra, neither highlighting alone nor combined with division of labor helped. In particular:

- Under the both treatments, none of the workers correctly identified all traps and clues, and 72% of labelers provided the answers consistent with highlights. As in case with Kitz, the excessive highlighting did not influence the workers' accuracy. For Andorra, 45% of texts contained excessive highlighting.

We further investigate the importance of highlighting alone. Table 10 contains the accuracy of Treatment 1 group when the highlighting was optional. In this table, we separated the accuracy of workers who did not use highlighting from those workers who used it.

Text	**Texts without highlights correct/total (%)**	**Highlighted texts correct/total (%)**
Gastein	13/16 (81%)	33/36 (92%)
Kitz	10/15 (67%)	29/37 (78%)
Andorra	6/20 (30%)	11/34 (32%)

Table 10: Accuracy of each text for the Treatment 2 group

Data suggest that **highlighting improved the accuracy in case of simple and moderate tasks asking to count relative objects in the text, and it had no effect in complex tasks**.

Events Task

Next, we describe the Events task. The binary version asks if it is possible for a tourist to visit all mentioned events during some period of time. The numerical task asks to provide the number of events that a tourist can visit. All three texts describe some events which take place outside the given time interval. Andorra text is designed as the simplest one - mentioned events provide multiple opportunities to visit all of them. In contrast, Gastein and Kitzbuel texts describe two events which take place only on Sundays. However, a worker is supposed to realize that the given timeframe includes several Sundays which make it possible to visit all the events. We made Kitzbuel text a bit more challenging by including a big number of places end events which a worker should consider.

The accuracy and mean absolute relative error of each text is shown in Tables 11 and 12. In addition, Table 11 contains the results of Treatment 3 group in which we decided to investigate how the quality of highlights influenced the labelers' accuracy. We created an additional Treatment 3 that was analogous to Treatment 2 with one difference. The second worker was provided with the text highlighted by a worker from the Treatment 1 group.

Text (difficulty)	Correct Answer	Control correct/total (%)	Treatment 1 correct/total (%)	Treatment 2 correct/total (%)	Treatment 3 correct/total (%)
Andorra (easy)	Yes	8/18 (44%)	13/16 (81%)	11/15 (73%)	12/19 (63%)
Gastein (normal)	Yes	7/8 (88%)	3/6 (50%)	16/20 (80%)	7/8 (88%)
Kitzbuel (difficult)	Yes	8/13 (62%)	4/11 (36%)	8/13 (62%)	3/8 (38%)

Table 11: Accuracy of each text for the binary Events task.

Text (difficulty)	Correct Answer	Control correct/total (%)	Treatment 1 correct/total (%)	Treatment 2 correct/total (%)
Andorra (easy)	10 events	0.327	0.288	0.230
Gastein (normal)	8 events	0.468	0.344	0.366
Kitzbuel (difficult)	11 events	0.273	0.338	0.168

Table 12: Mean absolute relative error of each text for the numerical Events task.

For Treatment 3 we didn't control for the quality of highlights before the experiment. After the experiment we checked the highlights and observed substantial differences. The majority of them skipped clues or highlighted irrelevant information. We see that Treatment 3 accuracy differs from Treatment 1 and Treatment 2 groups suggesting that **low quality highlights hinder the labelers' performance**.

Budget Task

Next, we provide the analysis of the Budget task. The binary version of it asks the participants if the minimum required budget to go on vacation to the ski-resort is less than 4500 euro. The numerical task requires submitting the exact budget. Kitzbuel text is the most difficult. If a worker missed the available discounts his budget would be very close to the boundary value leaving a small room for a mistake. The Gastein text also contains discounts however they are less crucial for the right conclusion. We did not include any discounts in the Andorra text, and designed it as the easies one. For this text, a worker could provide a wrong answer only if he made an arithmetic error, or included all optional items into his budget, or did not choose the cheapest available option (given that all options were listed in one paragraph). The accuracy for each text is shown in Tables 13 and 14. Table 13 also contains the results of Treatment 3 experiment, which was performed in the same way as in the Events task.

Text (difficulty)	Correct Answer	Control correct/total (%)	Treatment 1 correct/total (%)	Treatment 2 correct/total (%)	Treatment 3 correct/total (%)

Text (difficulty)	Correct Answer	Control correct/total (%)	Treatment 1 correct/total (%)	Treatment 2 correct/total (%)	Treatment 3 correct/total (%)
Andorra (easy)	Yes	4/13 (31%)	12/17 (71%)	9/14 (64%)	10/15 (67%)
Gastein (normal)	Yes	8/10 (80%)	9/14 (64%)	14/16 (88%)	16/20 (80%)
Kitzbuel (difficult)	Yes	7/11 (64%)	13/13 (100%)	11/15 (73%)	14/17 (82%)

Table 13: Accuracy of each text for the binary Budget task

Text (difficulty)	Correct Answer	Control correct/total (%)	Treatment 1 correct/total (%)	Treatment 2 correct/total (%)
Andorra (easy)	4370 euro	0.350	0.270	0.320
Gastein (normal)	4320 euro	0.277	0.264	0.321
Kitzbuel (difficult)	4300 euro	0.237	0.266	0.229

Table 14: Mean absolute relative error of each text for the numerical Budget task.

As before in Treatment 3 we got similar mixed results as in Events task.

Conclusion

Concluding our findings, for labor division we got statistically significant results in time reduction in the last stage of the task. Although the overall task completion took longer. Annotators took more time to prepare their results for labelers. However, labelers performed additional text analysis besides solely relying on the annotators' highlights.

Another observation is that text highlights may not be the best medium to transfer the result of work from one stage to another. They functioned well for the task asking to count relative objects in the text (Famose task) but rather impeded the workers in other types of tasks. Moreover, bad highlights worsen the labelers' accuracy.

We analysed surveys filled in by workers after they completed our tasks. We didn't identify any patterns in the crowd source population that would help to distinguish between workers who benefited from highlights or were hindered by them.

4 References

1. Jeffrey M. Rzeszotarski, Aniket Kittur. Instrumenting the Crowd: Using Implicit Behavioral Measures to Predict Task Performance, 2011.
2. Gondy Leroy, James E, Obay Mouradi, David Kauchak, Melissa L. Just. Improving Perceived and Actual Text Difficulty for Health Information Consumers using Semi-Automated Methods, 2012.

3. Robert F. Lorch, Jr. Text-Signaling Devices and Their Effects, 1989.
4. Kenneth E. Bell and John E. Limber. Reading Skill, Textbook Marking, and Course Performance, 2009.
5. Samuel Dodson, Luanne Freund, Rick Kopak. Do Highlights Affect Comprehension? Lessons from a User Study, 2017.
6. Ed H. Chi, Michelle Gumbrecht, and Lichan Hong, Visual Foraging of Highlighted Text: An Eye-Tracking Study, 2007.
7. Rosta Farzan and Peter Brusilovsky, Social Navigation Support for Information Seeking: If You Build It, Will They Come?, 2009.
8. Silvers, Vicki L, and David S Kreiner. The effects of pre-existing inappropriate highlighting on reading Comprehension, 1997.
9. Kristie Fisher, Scott Counts, Aniket Kittur. Distributed Sensemaking: Improving Sensemaking by Leveraging the Efforts of Previous Users, 2012.
10. Shneiderman, B. Designing trust into online experiences, 2000.
11. Rosta Farzan, A STUDY OF SOCIAL NAVIGATION SUPPORT UNDER DIFFERENT SITUATIONAL AND PERSONAL FACTORS, 2011.
12. Ria Mae Borromeo, Thomas Laurent, Motomichi Toyama, The Influence of Crowd Type and Task Complexity on Crowdsourced Work Quality, 2016
13. Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, Katrina Panovich, Soylent: A word processor with a crowd inside. ACM, 2010.
14. Joy Kim, Sarah Sterman, Allegra Argent Beal Cohen, Michael S. Bernstein, Mechanical Novel: Crowdsourcing Complex Work through Reflection and Revision, 2017
15. Lixiu Yu, Jeffrey V. Nickerson, Cooks or Cobblers? Crowd Creativity through Combination, 2011
16. Greg Little, Lydia B. Chilton, Max Goldman, Robert C. Miller, Exploring Iterative and Parallel Human Computation Processes, 2010.
17. Alexander J. Quinn, Benjamin B. Bederson, A Taxonomy of Distributed Human Computation, 2010.
18. Chris Callison-Burch and Mark Dredze, Creating Speech and Language Data With Amazon's Mechanical Turk, 2010.
19. D. E. Difallah et al. The dynamics of micro-task crowdsourcing: The case of amazon mturk. In WWW, pages 238–247, 2015.
20. Ayush Jain, Akash Das Sarma, Aditya Parameswaran, Jennifer Widom, Understanding Workers, Developing Effective Tasks, and Enhancing Marketplace Dynamics: A Study of a Large Crowdsourcing Marketplace, 2017.
21. J. Vuurens, A. P. de Vries, and C. Eickhoff. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval, 2011.