Just the right mood for HIT!

Analyzing the role of worker moods in conversational microtask crowdsourcing

Qiu, Sihang; Gadiraju, Ujwal; Bozzon, Alessandro

DOI
10.1007/978-3-030-50578-3_26

Publication date
2020

Document Version
Final published version

Published in
Web Engineering - 20th International Conference, ICWE 2020, Proceedings

Just the Right Mood for HIT!

Analyzing the Role of Worker Moods in Conversational Microtask Crowdsourcing

Sihang Qiu(✉), Ujwal Gadiraju, and Alessandro Bozzon

Web Information Systems Group, Delft University of Technology, Delft, Netherlands
{s.qiu-1,u.k.gadiraju-1,a.bozzon}@tudelft.nl

Abstract. Conversational agents are playing an increasingly important role in providing users with natural communication environments, improving outcomes in a variety of domains in human-computer interaction. Crowdsourcing marketplaces are simultaneously flourishing, and it has never been easier to acquire large-scale human input from online workers. Recent works have revealed the potential of conversational interfaces in improving worker engagement and satisfaction. At the same time, worker moods have been shown to have significant effects on quality related outcomes. Little is known about the role of worker moods in shaping work in conversational microtask crowdsourcing. In this paper, we conducted a crowdsourcing study addressing 600 unique online workers, to investigate the role that worker moods play in conversational microtask crowdsourcing. We also explore whether suitable conversational styles of the agent can affect the performance of workers in different moods. Our results show that workers in a pleasant mood tend to produce significantly higher quality results (over 20%), exhibit greater engagement (an increase by around 19%) and report a lower cognitive load (by over 12%), and a suitable conversational style can have a significant impact on workers in different moods. Our findings advance the current understanding of conversational microtask crowdsourcing and have important implications on designing future conversational crowdsourcing systems.

Keywords: Crowdsourcing · Conversational agent · Conversational style · Worker moods · Worker performance · Moods

1 Introduction

Microtask crowdsourcing is widely being used to gather human input in decomposed tasks called HITs (human intelligence tasks) [12]. Crowdsourcing HITs have been used for a variety of purposes – to build ground truths, understand human behavior, evaluate systems, among others [2,17,30]. Most of the popular commercial microtasking platforms (such as Amazon Mechanical Turk and FigureEight) provide workers with traditional web interfaces for task consumption

M. Bielikova et al. (Eds.): ICWE 2020, LNCS 12128, pp. 381–396, 2020.
https://doi.org/10.1007/978-3-030-50578-3_26

and execution. However, engaging workers in large batches of HITs is challenging. Task abandonment and drop-out effects are commonly observed in microtask marketplaces due to fatigue, boredom or other task-related factors [8].

Conversational interfaces have been argued to have advantages over traditional graphical user interfaces due to having a more human-like interaction [20]. Moreover, recent work has shown that conversational interfaces can be used to improve worker engagement and satisfaction in microtask crowdsourcing [9,18]. Worker *moods* are known to influence the quality of work in the workplace [26], including online microtasking platform where microtasks are executed using traditional web interfaces [28,31]. For example, workers in pleasant moods were found to significantly outperform those in unpleasant moods in a series of information finding HITs [6]. There is a limited understanding however, of how moods of workers interact with conversational interfaces in shaping the quality of their work. An unexplored opportunity to improve conversational microtasking further, lies in analyzing the potential impact of conversational styles [25] of agents on quality related outcomes of workers in different moods. Psychologists and linguists have found that conversational styles play an important role in communication [15,24,25]. Our recent study has investigated whether adapting and personalizing the conversational style of an agent to that of a worker can improve the quality of work [23]. We aim to fill this knowledge gap by addressing the following research questions:

RQ1: *How do worker moods affect their performance, engagement and cognitive load in conversational microtask crowdsourcing?*
RQ2: *How does the conversational style of a conversational agent affect the performance of workers in different moods?*

In this paper, we designed and implemented a conversational interface with different conversational styles that supports workers in the execution of HITs. We carried out a crowdsourcing study with 600 unique workers, across four types of tasks and three different interfaces ($3 \times 4 = 12$ experimental conditions in total). To answer **RQ1**, we evaluate the performance of workers, their engagement (using the User Engagement Scale-*UES*) and cognitive load (NASA-TLX) across different tasks. Results reveal that workers in a pleasant mood tend to produce significantly higher quality results (over 20% improvement), exhibit greater engagement (over 18% improvement) and report a lower cognitive load (a decrease by nearly 13%). To address **RQ2**, we considered three different interfaces (traditional web interface, and conversational interfaces with two conversational styles). Results demonstrate that a suitable conversational style can have a significant impact on workers in terms of their engagement and cognitive task load.

2 Related Work

2.1 Conversational Agents and Crowdsourcing

Conversational agents have been widely used in crowdsourcing workflows. Most studies have used conversational agents with an aim to train natural language

understanding and processing models [14]. Another popular application is the usage of conversational agents to connect users with crowd-powered Q&A systems. Such conversational agents act like a representative of the crowd, working for aggregating and conveying information from the crowd to the user. Lasecki et al. designed a conversational agent named Chorus, to help users acquire general knowledge from the crowd [16]. Huang et al. designed a series of conversational systems that improve the effectiveness of collaborative work done by workers [10,11]. In contrast, Curious Cat was designed for acquiring knowledge from users [1]. In this paper, we design and implement a conversational agent that is fully functional on an HTML-based webpage, and supports the execution of HITs.

2.2 Worker Moods in Crowdsourcing

Prior studies have established that worker moods in real-life can affect their task performance; workers in a happy mood were found to exhibit a better performance than those who were less happy [27,29]. Others have shown that worker moods can also impact task execution time [19]. Recent work in the context of online crowdsourcing has revealed the relationship between worker moods and crowdsourcing task performance [31], where moods were measured using the Pick-A-Mood instrument [3]. Statistical tests indicated that worker moods had significant effects on their engagement. Based on these findings, others analyzed the impact of worker moods in struggling web search tasks [6]. Due to the evident impact of worker moods on quality related task outcomes on traditional web interfaces, in this paper we analyze how worker moods interact with conversational interfaces to shape work quality.

3 Method

3.1 Workflow and Task Design

The entire task execution process across different conditions consists of four main stages: self-reported mood (Pick-A-Mood), a short demographic survey, the crowdsourcing HITs, and a post-task survey, as illustrated in Fig. 1.

1) *Pick-A-Mood.* Workers are first asked to self-report their moods using the Pick-A-Mood instrument shown in Fig. 2. Nine moods are presented, and can be grouped into three categories, which are **pleasant** moods (A: *cheerful*, B: *excited*, H: *relaxed* and G: *calm*), **unpleasant**-moods (C: *tense*, D: *irritated*, E: *sad* and F: *bored*) and a **neutral** mood (I).
2) *Demographic Survey.* Next, workers are asked to respond to simple background questions pertaining to their gender, age, ethnicity, educational background, and sources of income.
3) *Crowdsourcing HIT Design.* The actual crowdsourcing HITs are executed on either the conversational interface or the traditional web interface as per the experimental condition. The microtasks batch has 5 mandatory HITs and

45 optional HITs. Workers must complete the 5 mandatory HITs to proceed to the next stage. On completing the mandatory HITs in the conversational interface condition(s), the agent asks the workers if they want to continue on and complete more HITs. In case of the traditional web interface condition(s), workers can click a button named '`I want to answer more questions`' to complete more optional HITs.

4) *Post-task Survey.* The last stage of the workflow presents workers with a survey, to gather the worker's perception about the HITs completed. Workers are first asked to complete the User Engagement Scale Short Form [21,22] (UES-SF). Within this, 12 questions need to be answered by adjusting the slider bar ranging from "*1: Strongly Disagree*" to "*7: Strongly Agree*". O'Brien designed the UES for systematically measuring user engagement through self-assessment [21], and later developed the short form of UES (UES-SF) to be suitable for time-sensitive contexts [22]. Next, workers are asked to complete the NASA Task Load Index (NASA-TLX) questionnaire[1], which includes six questions corresponding to different kinds of cognitive task load (ranging from "*0: Very Low*" to "*100: Very High*").

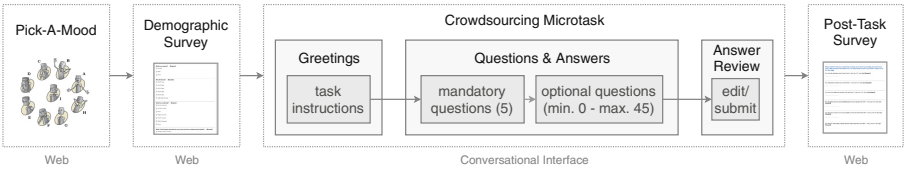


Fig. 1. Crowdsourcing microtask workflow in the conversational interface conditions.

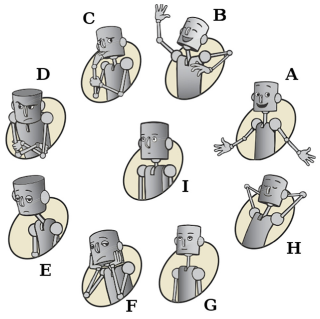


Fig. 2. Pick-A-Mood scale to measure the self-reported mood of crowd workers in different conditions.

[1] https://humansystems.arc.nasa.gov/groups/TLX/.

3.2 Conversational Interface

To support the execution of HITs on a conversational interface, we incorporate the following aspects.

1) *Greetings.* Drawing from the essential structure of conversation, the conversational interaction begins with greetings. The goal here is to let workers familiarize themselves with the conversational interface. Next, the conversational interface then helps workers understand how to execute HITs by introducing the task instructions using dialogues.
2) *Questions & Answers.* The conversational interface asks questions to workers, and workers can answer these questions by either typing answers or using provided UI (user interface) elements.
3) *Answer Review.* On the traditional web interface, a worker can easily go back to a question and edit its answer. To realize this affordance in the conversational interface, workers are provided with the opportunity to review and edit their answers if needed, before submitting the HITs.

The user interfaces of most common crowdsourcing platforms mainly support HTML/CSS and Javascript. To make sure the conversational interface can be directly embedded into such platforms, the conversational interface is developed based on a HTML/Javascript chatbot project `chat-bubble`[2]. This allows us to avoid redirecting workers to an external chatting or messaging application.

The conversational interface supports two modes of input– free text and multiple choices, since these two types of input can enable workers to effectively provide judgments for most popular crowdsourcing task types [5]. As shown in Fig. 3, `bubble-like buttons` and `textarea` (at the bottom of UI) are used for supporting the input modes of multiple choice selection and free text entry respectively. For HITs that need special functions (for example, drawing bounding boxes), the input mode of the conversational interface can be ported from traditional web interfaces with little effort, as the conversational agent that we developed fully supports HTML elements.

3.3 Conversational Style

We also investigate whether a suitable conversational style of the conversational agent can affect the performance of workers in different moods. According to Deborah Tannen's seminal theory, conversational styles can be classified into two broad categories, namely *High Involvement* and *High Considerateness* [24]. A conversational style is actually the superimposition of multiple linguistic features and devices [25]. To this end, we selected features and devices that can be applied in our case to create conversation agents emulating High Involvement and High Considerateness conversational styles according to the design criteria from the previous work [23]. Selected features are shown in Table 1. Table 2 shows examples of how the conversational agent opens a conversation while emulating the two different conversational styles.

[2] https://github.com/dmitrizzle/chat-bubble.

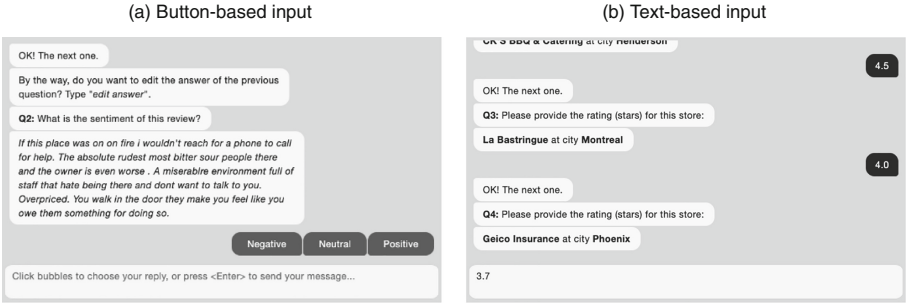


Fig. 3. Conversational interfaces for execution of HITs provide two input means: (a) buttons and (b) free text.

Table 1. Features of conversation used to design the conversational agents emulating different conversational styles [23].

Features	High-involvement	High-considerateness
Pace	Fast	Slow
Introduction of topics	Without hesitation	With hesitation
Use of syntax	Simple	Complex
Enthusiasm	Enthusiastic	Calm
Directness of content	Direct	Indirect
Use of questions	Frequent	Rare

4 Experiments and Setup

4.1 Experimental Design

In our experiments, we consider two data types (image and text) and two input types (free text and multiple choices), resulting in 4 HIT types (2 data types $\times$ 2 input types) - Information Finding (text data + free text input), Sentiment Analysis (text data + multiple choices), CAPTCHA Recognition (image data + free text input) and Image Classification (image data + multiple choices). The experiment is approved by the ethics committee of our university.

In **Information Finding (IF)** tasks, workers are asked to find and provide the rating (stars) of a given store from Google Maps. In **Sentiment Analysis (SA)** tasks, workers are asked to read given reviews of stores and determine the overall sentiment of the review. In **CAPTCHA Recognition (CR)** tasks, workers are asked to observe the image and determine which letters the image contains, and then provide the letters in the same order as they appear in the image. In **Image Classification (IC)** tasks, workers are asked to determine which animal the image contains.

We consider three distinct interfaces: **1) Traditional web interface (web)** where all the HITs are displayed and answered using traditional HTML elements;

Table 2. Examples of greetings with High-Involvement and High-Considerateness styles.

High involvement	High considerateness
— Hey! Can you help me with a task called Information Finding?	*— Thank you in advance for helping me with a task called Information Finding*
— You must complete this task within 30 min, otherwise I won't pay you	*— I think 30 min should be more than enough for you to finish*
— Here is the task instructions. Take a look!	*— I kindly ask you to have a look at the task instructions*

2) Conversational interface with High-Involvement style (Con+I), where the HITs are presented through an agent with a High-Involvement style; **3) Conversational interface with High-Considerateness style (Con+C)**, which is similar to Con+I, except that the agent converses with workers using a High-Considerateness style.

Thus, the four task types and three interfaces result in a cross-section of 12 experimental conditions. These 12 experimental conditions were published on Amazon Mechanical Turk (MTurk) as HIT batches in our experiments.

4.2 Evaluation Metrics

The evaluation metrics in our experiments are *output quality*, *worker engagement*, and *cognitive task load*.

Output quality is measured using the accuracy of workers. A worker's accuracy is calculated as the fraction of correct responses over the total number of responses provided by a worker. Here, we consider a HIT to be accurately completed if and only if the response is identical to the ground truth (case insensitive).

Worker engagement is measured using: 1) worker retention, quantified by the number of optional HITs completed (ranging from 0 to 45); and 2) the UES-SF scores ranging from 1 to 7. A higher UES-SF score indicates that the worker is relatively more engaged.

Cognitive task load is evaluated by unweighted NASA-TLX form, consisting of six questions. Workers are asked to give scores ranging from 0 to 100 to these questions. The final TLX score is the mean value of scores given to the six questions. The higher the TLX score is, the greater is the task load perceived by a worker.

4.3 Workers and Rewards

In our setup, each experimental condition consists of 50 HITs and we recruited 50 unique workers to participate and complete the workflow in each case. As a result, we acquired judgments from $12 \times 50 = 600$ unique workers.

After a worker provided a valid `task token` and successfully submitted the HITs on MTurk, the worker was immediately paid 0.5 USD, a fixed payment for successful submission. To reach an average hourly wage of 7.5 USD, we provided bonuses to workers according to the number of optional HITs that they completed. Workers working on image-based tasks (CAPTCHA Recognition and Image Classification) received 0.01 USD for each optional HIT, while workers working on text-based tasks (Information Finding and Sentiment Analysis) received 0.02 USD for each optional HIT.

4.4 Quality Control

Although MTurk allows task requesters to set a qualification type to prevent workers from executing tasks in multiple HIT batches, workers are still able to execute multiple HITs from a single batch. To ensure each worker at most submits once, we recorded unique worker IDs on our server using Javascript, to prevent repeated participation. To ensure reliability of results, validity of responses, and control for potential malicious activity [4,7], we restricted participation by using an MTurk qualification attribute, only allowing crowd workers whose HIT approval rates were greater than 95% to access our tasks.

5 Results

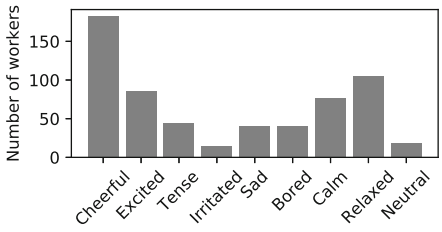
5.1 Worker Demographics

Of the unique 600 workers, 36.6% were female and 63.4% were male. The majority of workers were found to be Asian (46.37%), while 39.12% of workers were Caucasian. Most workers (89.2%) were under 45 years old, and education levels of most workers (74.5%) were higher than (or equal to) Bachelor's degree. In terms of source of income, 38.0% of the workers claimed MTurk was their primary source of income, while 55.4% of the workers worked on MTurk part-time and considered it as their secondary source of income. We publicly released all data (HITs deployed and responses from workers across the different experimental conditions) to facilitate further research for the benefit of the community[3].

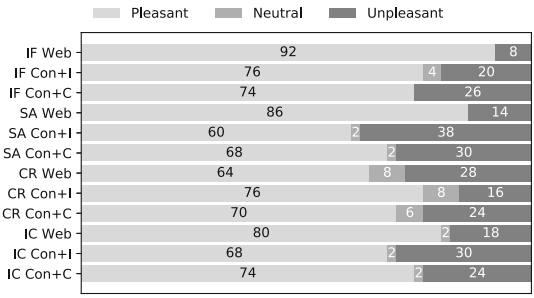
5.2 Distribution of Worker Moods

According to the results from the Pick-A-Mood instrument, 74.45% of workers reported to be in a pleasant mood, and 22.67% of workers reported unpleasant moods. Only 2.88% of workers reported to be in a neutral mood. As shown in Fig. 4(a), most workers reported to be in a cheerful mood. Consistent with prior findings in microtasking marketplaces [6,28,31], we found that a majority of workers were in pleasant moods.

[3] Companion page: https://sites.google.com/view/icwe2020mood.



(a)



(b)

Fig. 4. (a) Overall distribution of worker moods; (b) Percentages of workers in pleasant, neutral and unpleasant moods across different experimental conditions.

Figure 4(b) shows the distribution of worker moods across all experimental conditions, where IF, SA, CR and IC represent Information Finding, Sentiment Analysis, CAPTCHA Recognition, and Image Classification respectively. Web, Con+I and Con+C refer to the web interface, conversational interface with involvement-style and conversational interface with considerateness-style in each case. The mood distribution of workers within each experimental condition is similar to the overall mood distribution. Moreover, there were no workers who reported a neutral mood in web interface conditions of Information Finding and Sentiment Analysis tasks, and the conversational interface with High-Considerateness style of Information Finding (IF Web, IF Con+C and SA Web). Since there were only a few workers with a neutral mood who executed HITs across different experimental conditions, we excluded the workers in a neutral mood in our analysis presented further.

5.3 Worker Performance

We analyzed the performance of workers across different experimental conditions. Figure 5 shows the output accuracy of workers. Due to the relative ease of tasks, in case of image-based HITs (CAPTCHA Recognition and Image Classification), the output accuracy of workers is generally higher and more stable

across different interfaces and worker moods, compared to that in text-based HITs (Information Finding and Sentiment Analysis).

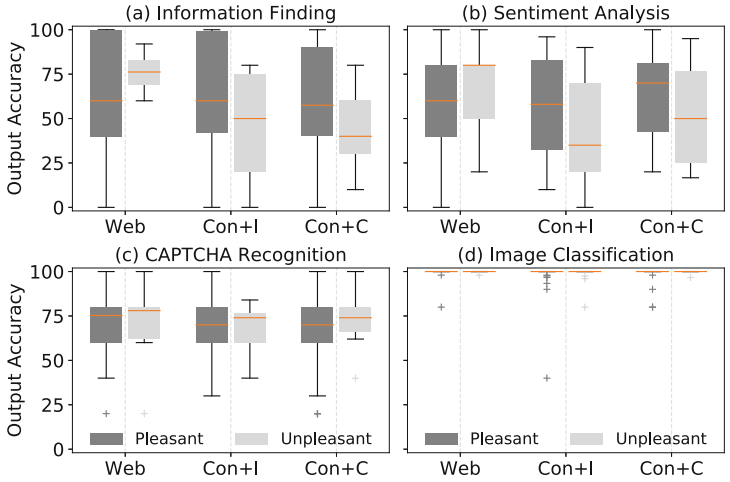


Fig. 5. Boxplots showing the output accuracy (unit: %) of workers in different moods, across different experimental conditions. Red lines in boxplots indicate the median value. (Color figure online)

To assess whether moods can affect worker performances in different interfaces, we conducted t-tests (two-tailed, $\alpha = 0.05$) to test the significance of pairwise differences between different interfaces within one conversational style. Results show that the performance of workers in unpleasant moods, using the conversational interface with High-Considerateness style (Con+C, $\mu = 43.1$, $\sigma = 23.0$) is significantly lower than those using the web interface (Web, $\mu = 76.1$, $\sigma = 11.6$) in Information Finding task (unpleasant, IF Con+C vs. IF Web, $p = 0.02$). In general, we found that the output quality corresponding to workers in unpleasant moods using conversational interfaces (both Con+I and Con+C) is generally lower than those using the traditional web interface on text-based tasks. This can intuitively be explained by the potential aversion of workers to engage with a conversation when in an unpleasant mood [13].

To investigate how workers with different moods perform under the same condition, we tested the statistical differences between the performance of workers across the two conversational styles using t-tests (two-tailed, $\alpha = 0.05$). Workers in pleasant moods performed significantly better than those in unpleasant moods, while using conversational interfaces with High-Involvement (pleasant $\mu \pm \sigma = 68.2 \pm 28.0$ vs. unpleasant $\mu \pm \sigma = 46.3 \pm 28.6$) and High-Considerateness styles (pleasant $\mu \pm \sigma = 63.3 \pm 29.8$ vs. unpleasant $\mu \pm \sigma = 43.1 \pm 23.0$) for executing Information Finding HITs (pleasant vs. unpleasant on IF Con+I and IF Con+C, $p = 0.031$ and $p = 0.033$ respectively). In general, our results suggest

that workers in pleasant moods exhibited a higher quality while using conversational interfaces, in comparison to workers in unpleasant moods.

5.4 Worker Engagement

Worker Retention. Fig. 6 shows the number of optional questions that workers answered across different task types, interfaces and moods. Since the number of optional HITs completed does not follow a normal distribution, we conducted Wilcoxon rank-sum tests (two-tailed, $\alpha = 0.05$) to test for statistical significace.

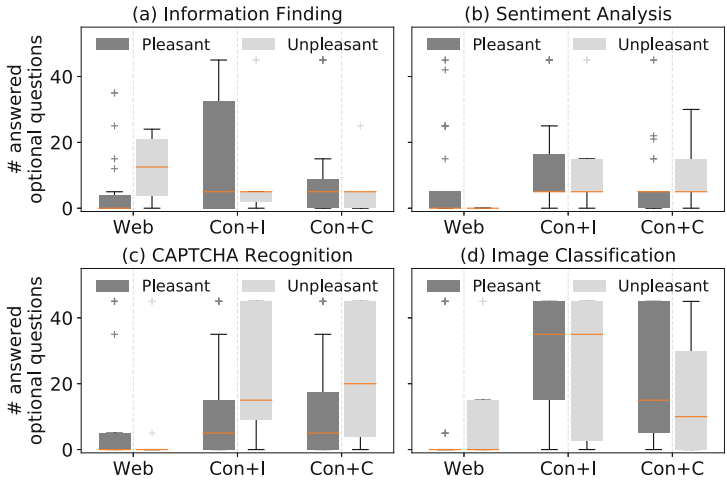


Fig. 6. Boxplots showing the number of optional HITs completed by workers in different moods across different experimental conditions. Red lines in the boxplots represent the median value. (Color figure online)

By comparing worker retention of different moods within each experimental condition, we found that the retention of workers in pleasant moods ($\mu = 7.2$, $\sigma = 10.7$) is significantly lower than that of workers in unpleasant moods ($\mu = 10.8$, $\sigma = 8.1$) using conversational interfaces with the Considerateness style for executing the Sentiment Analysis HITs (pleasant vs. unpleasant on SA Con+C, $p = 0.027$). This suggests that conversation interfaces with a particular conversational style can have the potential to improve worker retention based on the task type.

We found that workers in pleasant moods using conversational interfaces (both High Involvement and High Considerateness, Con+I and Con+C) answered significantly more optional HITs than workers in pleasant moods using traditional web interfaces across all four types of tasks (pleasant, all task types, $p < 0.05$). Workers in unpleasant moods also answered more optional HITs using conversational interfaces (both Con+I and Con+C) than those using web

interfaces in Sentiment Analysis and CAPTCHA recognition with significant differences (unpleasant, SA and CR, $p < 0.05$).

User Engagement Scale (UES-SF). We aggregated and analyzed the responses of workers in the post-task survey. Figure 7 depicts the UES-SF scores of workers across all types of tasks, interfaces and two different moods (pleasant vs. unpleasant). To understand the effect of worker moods on user engagement, t-tests (two tailed, $\alpha = 0.05$) are used to test the significance of differences.

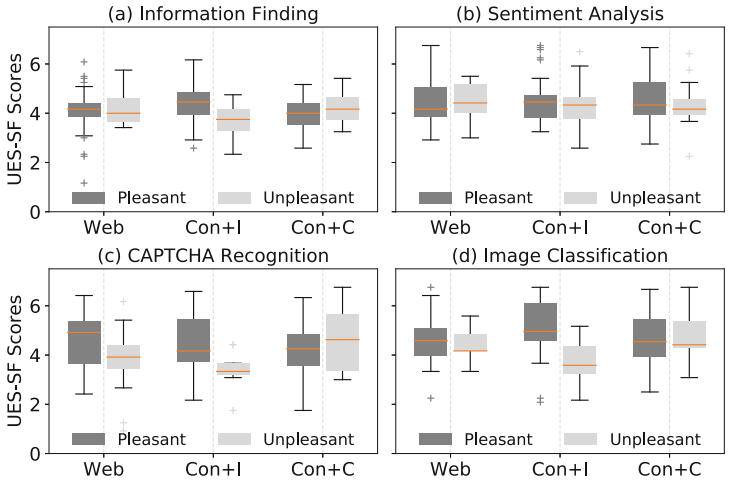


Fig. 7. UES-SF scores across different experimental conditions and worker moods. Red lines in the boxplots indicate the median value. (Color figure online)

Workers in pleasant moods reported significantly higher UES-SF scores than those in unpleasant moods on conversational interfaces with an involvement style (Con+I) for executing Information Finding (pleasant: $\mu = 4.4$, $\sigma = 0.8$ vs. unpleasant: $\mu = 3.7$, $\sigma = 0.7$), CAPTCHA Recognition (pleasant: $\mu = 4.4$, $\sigma = 1.1$ vs. unpleasant: $\mu = 3.4$, $\sigma = 0.8$), and Image Classification (pleasant: $\mu = 5.1$, $\sigma = 1.1$ vs. unpleasant: $\mu = 3.8$, $\sigma = 0.8$) HITs (pleasant vs. unpleasant on IF Con+I, CR Con+I and IC Con+I, $p = 0.02$, $p = 0.014$ and $p = 0.0001$ respectively).

UES-SF scores of workers in unpleasant moods using conversational interfaces with a considerateness style (Con+C) were significantly higher than those using conversational interfaces with an involvement style (Con+I) in CAPTCHA Recognition (Con+I $\mu \pm \sigma = 3.4 \pm 0.8$ vs. Con+C $\mu \pm \sigma = 4.6 \pm 1.3$) and Image Classification (Con+I $\mu \pm \sigma = 3.8 \pm 0.8$ vs. Con+C $\mu \pm \sigma = 4.7 \pm 1.0$) HITs (unpleasant, Con+I vs. Con+C in CR and IC, $p = 0.036$ and $p = 0.0125$ respectively). The High-Involvement conversational interface ($\mu = 4.4$, $\sigma = 0.8$) corresponds to significantly higher UES-SF scores than the High-Considerateness

conversational interface ($\mu = 3.9, \sigma = 0.7$) for workers in pleasant moods working on Information Finding HITs (pleasant, IF Con+I vs. IF Con+C, $p = 0.013$).

5.5 Cognitive Task Load

We also calculated the un-weighted NASA-TLX scores of all the workers participating in the crowdsourcing experiment. We use t-tests (two-tailed, $\alpha = 0.05$) to test the significance of differences between experimental conditions and worker moods (Fig. 8).

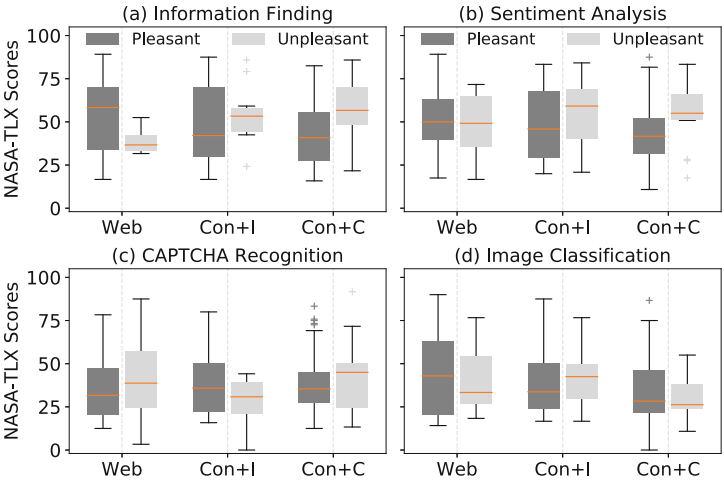


Fig. 8. NASA-TLX scores different experimental conditions and worker moods. Red lines in the boxplots indicate the median value. (Color figure online)

Workers in pleasant moods reported significantly lower NASA-TLX scores than workers in unpleasant moods in conversational interfaces with a High-Considerateness style (Con+C) for Information Finding (pleasant $\mu \pm \sigma = 42.8 \pm 19.1$ vs. unpleasant $\mu \pm \sigma = 55.4 \pm 18.1$) and Sentiment Analysis (pleasant $\mu \pm \sigma = 43.3 \pm 17.2$ vs. unpleasant: $\mu \pm \sigma = 54.9 \pm 18.1$) HITs (pleasant vs. unpleasant on IF Con+C and SA Con+C, $p = 0.046$ and $p = 0.041$ respectively). Thus, workers in pleasant moods perceived lesser cognitive task load in these conditions. Moreover, workers in pleasant moods also perceived less cognitive load while executing the Information Finding HITs on the conversational interface with a High-Considerateness style ($\mu = 42.8$, $\sigma = 19.1$), compared to the traditional web interface ($\mu = 53.5$, $\sigma = 21.1$) (pleasant, IF Con+C vs. IF Web, $p = 0.0200$).

6 Discussion

Implications. Our results clearly indicate that conversational interfaces for HIT execution can improve worker retention in general, irrespective of worker

moods. Statistical tests reveal the fact that pleasant workers were more engaged than unpleasant workers in general. This calls for the development and adoption of conversational interfaces for microtask crowdsourcing, and for methods to induce pleasant moods prior to HIT execution. Our results also suggest that conversational interfaces with a High-Considerateness style exhibit the potential to improve engagement of workers in unpleasant moods, while a High-Involvement style exhibits a potential to further engage workers in pleasant moods. In terms of cognitive task load, our findings show that workers in pleasant moods can perceive less task load than those in unpleasant moods while executing text-based HITs, especially when the conversational agent uses a High-Considerateness style. These findings present opportunities for task routing based on worker moods and by leveraging different conversational styles.

Caveats and Limitations. Despite the measures we took to ensure the reliability of responses of workers, as with any research that involves human subjects using self-reporting tools, a threat to the validity of our findings is the veracity of the self-reported moods of workers. However, the overall distribution of crowd worker moods are consistent with prior works that indicate a skew towards pleasant moods [6,31]. The mood distribution of workers is naturally unbalanced. It is however, not ethically sound to elicit unpleasant moods among workers to study the interaction between their moods and conversational styles of an agent.

7 Conclusions and Future Work

Through an experimental study in this paper, we explored how worker moods can affect their output quality, engagement and cognitive task load in conversational microtask crowdsourcing (**RQ1**). We also investigated how the conversational style of the conversational agent can affect the performance of workers in different moods (**RQ2**). We addressed **RQ1** by evaluating worker performance across different tasks. We addressed **RQ2** by comparing quality related outcomes between different interfaces (and conversational styles).

We found that workers in a pleasant mood generally exhibited a higher output quality (over 20% in the best case), higher user engagement (over 18%) and around 13% lesser cognitive task load. We also found strong evidence to suggest that a suitable conversational style can have a significant impact on worker performance under some specific conditions (such as the type of HIT). In the imminent future, we will explore the relationship between worker moods and their preferred conversational style.

References

1. Bradeško, L., Witbrock, M., Starc, J., Herga, Z., Grobelnik, M., Mladenić, D.: Curious cat-mobile, context-aware conversational crowdsourcing knowledge acquisition. ACM Transactions on Information Systems (TOIS) **35**(4) (2017). Article no. 33

2. Demartini, G., Difallah, D.E., Gadiraju, U., Catasta, M., et al.: An introduction to hybrid human-machine information systems. Found. Trends® Web Sci. **7**(1), 1–87 (2017)
3. Desmet, P.M., Vastenburg, M.H., Romero, N.: Mood measurement with Pick-A-Mood: review of current methods and design of a pictorial self-report scale. J. Des. Res. **14**(3), 241–279 (2016)
4. Eickhoff, C., de Vries, A.P.: Increasing cheat robustness of crowdsourcing tasks. Inf. Retrieval **16**(2), 121–137 (2013). https://doi.org/10.1007/s10791-011-9181-9
5. Gadiraju, U., Checco, A., Gupta, N., Demartini, G.: Modus operandi of crowd workers: the invisible role of microtask work environments. Proc. ACM Interact. Mob. Wearable Ubiquit. Technol. **1**(3) (2017). Article no. 49
6. Gadiraju, U., Demartini, G.: Understanding worker moods and reactions to rejection in crowdsourcing. In: Proceedings of the 30th ACM Conference on Hypertext and Social Media, HT 2019, pp. 211–220. ACM, New York (2019)
7. Gadiraju, U., Siehndel, P., Fetahu, B., Kawase, R.: Breaking bad: understanding behavior of crowd workers in categorization microtasks. In: Proceedings of the 26th ACM Conference on Hypertext & Social Media, pp. 33–38 (2015)
8. Han, L., et al.: All those wasted hours: on task abandonment in crowdsourcing. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pp. 321–329. ACM (2019)
9. Harms, J., Kucherbaev, P., Bozzon, A., Houben, G.: Approaches for dialog management in conversational agents. IEEE Internet Comput. **23**(2), 13–22 (2019)
10. Huang, T.H.K., Chang, J.C., Bigham, J.P.: Evorus: a crowd-powered conversational assistant built to automate itself over time. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, p. 295. ACM (2018)
11. Huang, T.H.K., Lasecki, W.S., Bigham, J.P.: Guardian: a crowd-powered spoken dialog system for web APIs. In: Third AAAI Conference on Human Computation and Crowdsourcing (2015)
12. Kittur, A., et al.: The future of crowd work. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work, pp. 1301–1318. ACM (2013)
13. Koch, A.S., Forgas, J.P., Matovic, D.: Can negative mood improve your conversation? Affective influences on conforming to Grice's communication norms. Eur. J. Soc. Psychol. **43**(5), 326–334 (2013)
14. Kucherbaev, P., Bozzon, A., Houben, G.J.: Human-aided bots. IEEE Internet Comput. **22**(6), 36–43 (2018)
15. Lakoff, R.T.: Stylistic strategies within a grammar of style. Ann. N. Y. Acad. Sci. **327**(1), 53–78 (1979)
16. Lasecki, W.S., Wesley, R., Nichols, J., Kulkarni, A., Allen, J.F., Bigham, J.P.: Chorus: a crowd-powered conversational assistant. In: Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology, pp. 151–162. ACM (2013)
17. Loni, B., Cheung, L.Y., Riegler, M., Bozzon, A., Gottlieb, L., Larson, M.: Fashion 10000: an enriched social image dataset for fashion and clothing. In: Proceedings of the 5th ACM Multimedia Systems Conference, pp. 41–46. ACM (2014)
18. Mavridis, P., Huang, O., Qiu, S., Gadiraju, U., Bozzon, A.: Chatterbox: conversational interfaces for microtask crowdsourcing. In: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, pp. 243–251. ACM (2019)
19. Miner, A.G., Glomb, T.M.: State mood, task performance, and behavior at work: a within-persons approach. Organ. Behav. Hum. Decis. Process. **112**(1), 43–57 (2010)

20. Moore, R.J., Arar, R., Ren, G.J., Szymanski, M.H.: Conversational UX design. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pp. 492–497. ACM (2017)
21. O'Brien, Heather: Theoretical perspectives on user engagement. In: O'Brien, Heather, Cairns, Paul (eds.) Why Engagement Matters, pp. 1–26. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-27446-1_1
22. O'Brien, H.L., Cairns, P., Hall, M.: A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. Int. J. Hum. Comput. Stud. **112**, 28–39 (2018)
23. Qiu, S., Gadiraju, U., Bozzon, A.: Improving worker engagement through conversational microtask crowdsourcing. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–12. ACM (2020)
24. Tannen, D.: Conversational style. In: Psycholinguistic Models of Production, pp. 251–267 (1987)
25. Tannen, D.: Conversational Style: Analyzing Talk Among Friends. Oxford University Press, Oxford (2005)
26. Totterdell, P., Niven, K.: Workplace Moods and Emotions: A Review of Research. Createspace Independent Publishing, Charleston (2014)
27. Wright, T.A., Cropanzano, R.: The happy/productive worker thesis revisited. In: Research in Personnel and Human Resources Management, pp. 269–307. Emerald Group Publishing Limited (2007)
28. Xu, L., Zhou, X., Gadiraju, U.: Revealing the role of user moods in struggling search tasks. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1249–1252. ACM (2019)
29. Zelenski, J.M., Murphy, S.A., Jenkins, D.A.: The happy-productive worker thesis revisited. J. Happiness Stud. **9**(4), 521–537 (2008). https://doi.org/10.1007/s10902-008-9087-4
30. Zhang, Z., Singh, J., Gadiraju, U., Anand, A.: Dissonance between human and machine understanding. Proc. ACM Hum.-Comput. Interact. **3**(CSCW) (2019). Article no. 56
31. Zhuang, M., Gadiraju, U.: In what mood are you today?: An analysis of crowd workers' mood, performance and engagement. In: Proceedings of the 11th ACM Conference on Web Science, WebSci 2019, Boston, MA, USA, 30 June–03 July 2019, pp. 373–382 (2019). https://doi.org/10.1145/3292522.3326010