Estimating Conversational Styles in Conversational Microtask Crowdsourcing

SIHANG QIU, Delft University of Technology, Netherlands
UJWAL GADIRAJU, Delft University of Technology, Netherlands
ALESSANDRO BOZZON, Delft University of Technology, Netherlands

Crowdsourcing marketplaces have provided a large number of opportunities for online workers to earn a living. To improve satisfaction and engagement of such workers, who are vital for the sustainability of the marketplaces, recent works have used conversational interfaces to support the execution of a variety of crowdsourcing tasks. The rationale behind using conversational interfaces stems from the potential engagement that conversation can stimulate. Prior works in psychology have also shown that '*conversational styles*' can play an important role in communication. There are unexplored opportunities to estimate a worker's conversational style with an end goal of improving worker satisfaction, engagement and quality. Addressing this knowledge gap, we investigate the role of conversational styles in conversational microtask crowdsourcing. To this end, we design a conversational interface which supports task execution, and we propose methods to estimate the conversational style of a worker. Our experimental setup was designed to empirically observe how conversational styles of workers relate with quality-related outcomes. Results show that even a naive supervised classifier can predict the conversation style with high accuracy (80%), and crowd workers with an *Involvement* conversational style provided a significantly higher output quality, exhibited a higher user engagement and perceived less cognitive task load in comparison to their counterparts. Our findings have important implications on task design with respect to improving worker performance and their engagement in microtask crowdsourcing.

CCS Concepts: • **Information systems** → **Chat**; **Crowdsourcing**; • **Human-centered computing** → **Empirical studies in HCI**.

Additional Key Words and Phrases: Conversational style; microtask crowdsourcing; work outcomes; user engagement; cognitive task load.

ACM Reference Format:
Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Estimating Conversational Styles in Conversational Microtask Crowdsourcing. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 32 (May 2020), 23 pages. https://doi.org/10.1145/3392837

1 INTRODUCTION

Crowdsourcing has become a primary means to gather human input for a variety of purposes: to build groundtruth, create datasets, evaluate systems, and to run human-centered experiments [3, 10, 61] among others. Currently, most online crowdsourcing tasks are executed on web platforms, like Amazon Mechanical Turk and Figure Eight, where workers usually view and complete tasks

Authors' addresses: Sihang Qiu, Delft University of Technology, Netherlands, s.qiu-1@tudelft.nl; Ujwal Gadiraju, Delft University of Technology, Netherlands, u.k.gadiraju-1@tudelft.nl; Alessandro Bozzon, Delft University of Technology, Netherlands, a.bozzon@tudelft.nl.

2573-0142/2020/5-ART32
https://doi.org/10.1145/3392837

using HTML-based interfaces. However, it is quite common that large batches of tasks suffer from worker drop-outs during the course of task execution [19, 20]. Many factors can cause task abandonment and make workers less engaged, such as the boredom, fatigue and inappropriate task difficulty levels [31, 36, 60].

To tackle this problem, researchers have introduced conversational agents into the realm of crowdsourcing. With *conversational microtask crowdsourcing* we refer to crowd work executed through a conversational agent that can provide online workers with a natural way to interact with a crowdsourcing system [4, 21, 25, 34]. Recent studies found that conversational microtask crowdsourcing can be used as an alternative to assist workers in task execution, reducing boredom [43], and producing similar output quality and task execution time, while leading to a high user satisfaction when compared to traditional web interfaces [37].

Previous works in the field of psychology have shown the important role that *conversational styles* have on inter-human communication [33, 49, 50]. Having been developed in the context of human conversations, the insights and conclusions of these works are not directly applicable to conversational microtasking, since the contrasting goal of workers is to optimally allocate their effort rather than being immersed in conversations. Previous work studied how specific linguistic strategies and language styles can affect work outcomes [27, 46]. In a recent study, we investigated whether different conversational styles of an agent can increase worker engagement [44]. However, to the best of our knowledge, current conversational agents (particularly for crowdsourcing) have only studied the effects of the conversational style of agents, rather than the conversational style of online users (i.e., workers in the context of microtask crowdsourcing). Understanding the role of workers' conversational styles in human computation can help us better adapt strategies to improve output quality and worker engagement, or better assist and guide workers in the training process. To this end, there is the need for novel methods for the classification of conversational styles in the context of microtask crowdsourcing. In this work, we will delve into the following research questions:

RQ1: *How can the conversational style of a crowd worker be reliably estimated?*

RQ2: *To what extent does the conversational style of crowd workers relate to their work outcomes, perceived engagement, and cognitive task load in different types of tasks?*

Original Contributions. We designed and implemented a conversational agent that supports crowdsourcing task execution and extraction of linguistic features from the text-based conversation between the user and the agent. We designed a coding scheme inspired by previous work [49, 50] and corresponding to conversational styles based on the five dimensions of linguistic devices that have been examined. Demonstrating the practical utility of our findings in this paper, we propose methods to predict the conversational style of users using either rich conversational data, or limited conversational data.

To evaluate our methods, we recruited 180 unique online crowd workers from Amazon Mechanical Turk (MTurk) and conducted experiments to investigate the feasibility of conversational style estimation for online crowdsourcing. We also analyzed the impact of conversational style on output quality, worker engagement (using worker retention and the User Engagement Scale), and perceived task load (using the NASA-TLX instrument). Our results show that we can predict the conversational style of workers using rich conversation data with a high accuracy (gradient boost: 80%), while we can also predict their conversational style using limited conversation data with an acceptable accuracy (gradient boost: 67%). Furthermore, our experimental findings revealed that workers with an *Involvement* conversational style have significantly higher output quality, higher user engagement and less cognitive task load while they are completing a high-difficulty

task, and have less task execution time in general. The findings have important implications on worker performance prediction, task scheduling and assignment in microtask crowdsourcing. To the best of our knowledge, this is the first work that explores the impact of conversational style on quality-related outcomes in conversational microtasking, and proposes methods to estimate the conversational style of users. To facilitate further research and for the benefit of the CSCW and HCI community, we publicly release our data and code[1].

The remainder of the paper is organized as follows: Section 2 discusses related work with regard to conversational microtask crowdsourcing and conversational styles. Section 3 depicts the workflow and structure of the conversational agent for microtask crowdsourcing, and our design of the conversational user interface. Section 4 describes our proposed approaches for conversational style estimation. Section 5 elaborates on our experimental design and configurations. In Section 6, we evaluate our methods and present results. In Section 7 and Section 8, we summarize the implications and limitations of our work, before drawing conclusions and setting precedents for future work.

2 RELATED WORK

2.1 Conversational Agents

Conversational interfaces have been argued to have advantages over traditional graphical user interfaces due to having a more human-like interaction [38]. Owing to this, conversational interfaces are on the rise in various domains of our everyday life and show great potential to expand [59]. Recent work in the CSCW and HCI community has investigated the experiences of people using conversational agents, understanding user needs and user satisfaction [8, 9, 35]. Other works have studied the scope of using conversational agents in specific domains. Vandenberghe introduced the concept of bot personas, which act as off-the-shelf users to allow design teams to interact with rich user data throughout the design process [56]. Others have studied the use of conversational agents in the domains of complex search [2, 28, 57] or food tracking [17]. These works have shown that conversational agents can improve user experiences and have highlighted the need to further investigate the use of conversational agents in different scenarios. Building on the findings of prior works, in this paper we investigate the potential of estimating worker's conversational styles to improve quality related outcomes in conversational microtasking.

2.2 Conversational Microtask Crowdsourcing

Prior research has combined crowdsourcing and the conversational agent for training the dialogue manager or natural language processing component [32]. Lasecki et al. designed and developed Chorus, a conversational assistant able to assist users with general knowledge tasks [34]. Conversations with Chorus are powered by workers who propose responses in the background, encouraged by a game-theoretic incentive scheme. Workers can see the working memory (chat history) and vote on candidate responses on a web-based worker interface. Based on Chorus, an improved conversational assistant named Evorus was proposed. It can reduce the effort of workers by partially automating the voting process [21]. The same authors also developed a crowdsourced system called Guardian, which enables both expert and non-expert workers to collaboratively translate Web APIs into a dialogue system format [22]. Conversational microtask crowdsourcing is also deployed on social network platforms, combing with messaging applications, such as Facebook and Twitter. Savage et al. designed a platform named Botivist based on Twitter, engaging volunteers to action by using different strategies [46]. A previous work based on Facebook Messenger used a Chatbot to connect learners and experts, for providing experts' feedback to improve learners' work [54].

[1] https://qiusihang.github.io/convsty/index.html

A conversational agent called Curious Cat was proposed to combine the crowdsourcing approach from a different perspective [4]. While most crowdsourced conversational agents provide information to users according to their requests, the Curious Cat was designed as a knowledge acquisition tool, which actively asked data from users. More recently, Mavridis et al. studied the feasibility of using a conversational interface as an alternative to traditional web interfaces on crowdsourcing platforms [37]. In our study, we propose a conversational agent that serves in a conversational interface, and facilitates the execution of different types of popular crowdsourcing tasks.

2.3 Conversational Style and Work Outcomes

The earliest systematic analysis of conversational style was performed by Lakoff [33]. She classified the stylistic strategies people used in everyday conversation into four categories, namely, *clarity*, *distance*, *deference*, and *camaraderie*. Lakoff found that speakers tend to use the camaraderie strategy when they want to be friendly, and use clarity strategy when they want to have the least relationship with another. Speakers can use a combination of different strategies in practice. Based on that, Tannen proposed a classification of conversational style that distributes speakers on a continuum from *High Considerateness* to *High Involvement*. She also concluded important features and linguistic devices helping in classification of speakers' conversational styles [49, 50].

In terms of the usage of conversational style in human-computer interaction, Shamekhi et al. analyzed the preferred conversational style of users for a virtual agent [48]. They extrapolated the conversational style by indirectly asking users about their attitudes during a conversation. Thomas et al. analyzed styles of the information-seeking conversation from the MISC dataset [53] by using some measurable properties to represent features defined by Tannen, such as pronoun use, speech rate, pitch/loudness variation, and so on [52].

Conversational styles can also affect work outcomes and worker performance. Using Botivist [46], the authors analyzed how strategies (corresponding to different language styles) could potentially affect the outcome. Previous work evaluated the impact of linguistic style matching (LSM [16]) on team performance in long-term tasks [39]. Tausczik et al. designed a real-time language feedback system to test the work outcomes of student groups by monitoring communication patterns [51]. Many prior studies have used automated methods to predict and analyze age, gender and personality based on linguistic features [1, 6, 40, 47]. A recent study compared the impacts of conversational styles on online surveys [27]. Authors defined two styles, "casual" and "formal" respectively, and then applied these two styles on web platform and chatbot. They concluded that a chatbot could play a role as a human in surveys if an appropriate conversational style is used.

Previous works however, are not directly applicable in microtask crowdsourcing. In our study, we perform analysis under Tannen's classification and definition of the conversational style. We propose conversational style estimation methods based on features and linguistic devices from High-Involvement and High-Considerateness styles. We evaluate how conversational styles of crowd workers affect work outcomes by measuring the worker accuracy, User Engagement Scale Short Form (UES-SF)[42], and NASA Task Load Index (NASA-TLX) in two popular types of crowdsourcing tasks – Information Finding and Image Transcription tasks.

3 CONVERSATIONAL MICROTASK CROWDSOURCING

Advances in microtask crowdsourcing have enabled the possibility of accomplishing even complex tasks by relying on crowd workers. Tasks such as speech transcription and copy-editing can be easily accomplished using microtask marketplaces, and researchers have found that even complex and expert tasks such as writing, product design, or translation may be amenable to novice crowd workers with appropriate process design and technological support [26, 29, 30, 55]. A notable feature of the interaction between crowdsourcing platforms and workers in the majority of prior

work, is the use of traditional web-based GUIs to communicate with workers, transmit instructions and gather responses thereafter. In the recently introduced notion of 'conversational microtask crowdsourcing' however, a conversational agent interfaces online workers with the crowdsourcing platform, facilitating task execution and task completion (cf. Mavridis et al. [37]).

3.1 Conversational Agent Design

The traditional web-based user interface of a HIT (human intelligence task) typically comprises of two parts: task instructions and microtasks. Workers are asked to first read instructions and then execute microtasks accordingly. To realize interaction comparable to web-based interfaces, we designed a text-based conversational agent composed of four phases [37], as described below.

Task instructions. Simulating the essence of natural conversation, the conversational agent begins the conversation with greetings, and then presents task instructions via a dialogue with the workers. The goal of this phase is to let workers familiarize themselves with the conversational agent and help them understand how to complete the crowdsourcing tasks.

Questions & Answers. The conversational agent asks questions (each question corresponds to a microtask) to workers, and workers can provide responses to microtasks by either typing answers or using UI elements. The conversational flow of this phase is specifically designed according to the task type.

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 (by typing "edit answer" to enable the selection of an answer to modify), before submitting the microtasks.

Reward. After reviewing the answers, workers enter the final stage where they can submit their answers and receive `Task Tokens` that they need to use to claim their rewards.

3.2 Text-based Conversational Interface

Popular crowdsourcing platforms (such as Amazon Mechanical Turk, Figure Eight, or Yandex Toloka) offer web interfaces based on standard technology like HTML, CSS and Javascript. To avoid the need for installing a messaging application – e.g. Telegram, or Whatsapp, where conversational agents are usually deployed, we designed and implemented a text-based conversational interface in HTML/CSS/Javascript, thus enabling easy integration with existing platforms and access to the available crowd workers.

The conversational interface supports any data source that is supported by HTML5, including text, image, audio and video. Therefore, most common task types such image classification, sentiment analysis, information finding, object recognition, speech transcription can all be implemented. Our design provides workers with two means to answer microtasks, as shown in Figure 1. Workers can either type in the textarea (a) or click a button (b) to send their responses. These two input types can support most of popular task types [13]. For some tasks that need special functions, UI elements from traditional web pages (e.g. slide bars, drawing tools, etc.) can also be included. In addition, the conversational interface records all the activities of the worker (including all keypress events with timestamps). The code for our text-based conversational interface is available to the community to facilitate further research.

4 CONVERSATIONAL STYLE ESTIMATION

Conversational interfaces can be an effective means for task execution [37]. Emulating particular conversational styles suitable to given contexts, or aligning the conversational style of an agent to the preferred style of workers, may help to improve worker engagement, satisfaction, and even output quality. To enable further research in this direction, we first need a reliable method

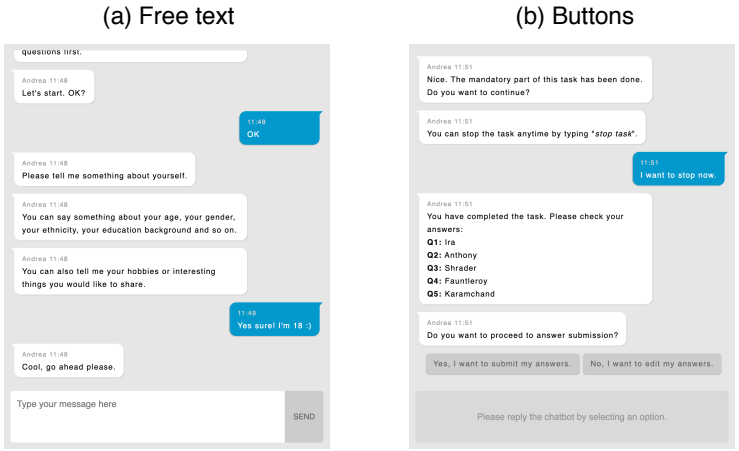


Fig. 1. Two interaction types of the conversational interface.

to estimate the conversational style of a worker. In this section, we first introduce background work on conversational styles (Section 4.1), and present a novel coding scheme designed to label and estimate conversation style of workers in conversational microtasking (Section 4.2). We then propose two approaches for conversational style prediction, based on rich and limited conversation data from workers.

4.1 High Involvement and High Considerateness

In this work we draw from Deborah Tannen's theory of conversational style [49, 50]. In Tannen's theory, conversational style can be classified broadly into two categories: *High Involvement* and *High Considerateness*. The High-Involvement style is described as follows: *"When in doubt, talk. Ask questions. Talk fast, loud, soon. Overlap. Show enthusiasm. Prefer personal topics, and so on"*. In contrast, she characterized the High-Considerateness style as follows: *"Allow longer pauses. Hesitate. Don't impose one's topics, ideas, personal information. Use moderate paralinguistic effects, and so on"*.

As per Tannen's theory, conversational styles emerge through the combined use of different linguistic devices. At the end of her book, Tannen identifies nine *dimensions* of linguistic devices that are related to conversational styles: *Personal focus of topic*, *Paralinguistic features*, *Enthusiasm*, *Use of questions*, *Pacing*, *Use of repetition*, *Topic cohesion*, *Tolerance of silence*, and *Laughter* [50]. She then presented an example of how conversational styles could be estimated: she created four continua (which could be extended, if needed) corresponding to four linguistic devices mapped on the nine dimensions above: humor, pace, percentage of narrative turns, and narratives [50]. These continua are superimposed upon one another to get an overall continuum. Participants who receive high scores on the overall continuum are classified as High-Involvement speakers, while those who receive low scores are classified High-Considerateness speakers.

4.2 Coding Scheme of Conversational Style

While providing a conceptual framework for the definition and characterisation of conversational styles, Tannen's theory is not directly applicable to conversational microtasking. Tannen's work was developed (and tested) in the context of human conversations, which are typically long and articulated. In conversational microtasking, devices like "humor" and "the percentage of narrative turns" are clearly at odds with the need for workers to optimally allocate their effort. Moreover,

Tannen's continua-based method for conversational style estimation does not have specific criteria to guide readers to distribute speakers on continua. For these reasons, a novel coding scheme for systematically classifying the conversational style is required, to enable the classification of coding styles, and guide the creation of ground truth data for conversation style estimation. This coding scheme builds upon a subset of the linguistic dimensions listed in the previous section. We exclude *Paralinguistic features*, *Use of repetition* and *Laughter*.

Several paralinguistic features, such as pitch shifting and voice quality, are usually absent in text-based chat. Repetition is fairly often used in voice chats, but it is absent in text chats [24], which can be explained by the ability for users to see previous utterances on the computer screen. Finally, we ignore the effects of text-based laughter as previous work suggests that nonlinguistic (nonverbal) cues (such as smile, laughter) are mostly absent from text-based conversations [45].

Table 1. Coding scheme for conversational style.

Dimension	**Score**	**Criteria**
1) Personal focus of topic	1	The worker prefers responding to the questions with personal opinions or personal anecdotes. For example, the worker uses first-person pronouns and phrases such as "I think", "I like", "my experience".
	-1	The worker prefers responding to questions by using objective descriptions. For example, using impersonal phrases such as "it is".
2) Enthusiasm	1	The worker demonstrates a willingness to converse with the conversational agent. For example, by responding positively to questions from the agent that would prolong the conversation.
	-1	The worker appears to be disinterested in the conversation with the agent. For example, by constantly seeking to end the conversation and responding with "no more", "nothing else", or similar phrases.
3) Pacing	1	Calculate the mean *pace* (typing rate) of all the workers. The score of the worker whose mean *pace* $\geq$ *median* is 1 (relatively faster pace).
	-1	Calculate the mean *pace* of all the workers. The score of the worker whose mean *pace* $<$ *median* is -1 (relatively slower pace).
4) Tolerance of silence	1	Calculate the mean *percentage of self-editing* (fractions of deleted characters among all the typed characters) of all the workers. The score of the worker whose mean *percentage of self-editing* $<$ *median* is 1.
	-1	Calculate the mean *percentage of self-editing* of all the workers. The score of the worker whose mean *percentage of self-editing* $\geq$ *median* is -1.
5) Topic cohesion & Use of questions	1	The worker prefers to express opinions directly linked to the topic or asks questions when in doubt.
	-1	The worker deviates from the topic without asking questions, but by responding respectfully to the conversational agent when in doubt.

We include *Tolerance of silence* in the coding scheme, i.e. hesitation and silence occurring in conversations, but with some adaptation. In text-based chat, we measure tolerance of silence through editing actions (i.e., when users edit a message before it is sent). We calculate the percentage of deleted keys among all the keys pressed by the worker. The higher the percentage is, the more hesitation the worker has, implying longer silence during the conversation.

In our study, *Topic cohesion* refers to whether the answers that workers give to pre-defined questions (described later in Section 5) are topically coherent, and well linked. In some cases

however, workers might directly ask questions to the conversational agent, referring to *4) Use of questions*, or express apologies to explain that they can not answer. Such questions or statements naturally deviate from the topic at hand. Therefore, we combine these two dimensions together as one factor in the coding scheme. The resulting set of dimensions used to systematically analyze conversation styles are summarised in Table 1, and they include: *1) Personal focus of topic*, *2) Enthusiasm*, *3) Pacing*, *4) Tolerance of silence*, and *5) Topic cohesion & Use of questions*.

Each dimension is quantified using a binary score (either -1 or 1). A final *score* is used to classify a conversation style as either *Involvement* or *Considerateness*. The score is calculated as a sum of scores corresponding to all the five dimensions. If final *score* is greater than 0, the conversational style of a worker is classified as *Involvement*. If the final *score* is less than 0, the conversational style of a worker is classified as *Considerateness*.

The coding scheme can be used to label ground truth data pertaining to conversational styles. To make the ground truth reliable, the coding process is carried out by multiple coders independently. Coders in the group independently score all the dimensions. The cases with disagreement are then resolved through manual discussion and disambiguation. The reliability of the coding process is measured by using Fleiss' Kappa [12].

4.3 Conversational Style Prediction

This section describes two methods for conversational style prediction: one based on "rich conversation data" obtained from workers before microtask execution through a "pre-task conversation", and another based on "limited conversation data" obtained from task execution and a short "self-assessment". Figure 3 describes the organisation of the envisioned conversational crowdsourcing task, where rich and limited conversation data are collected. Hereafter, we will refer to these methods as **Prediction with rich data** and **Prediction with limited data** respectively.

Table 2. Features used for conversational style prediction.

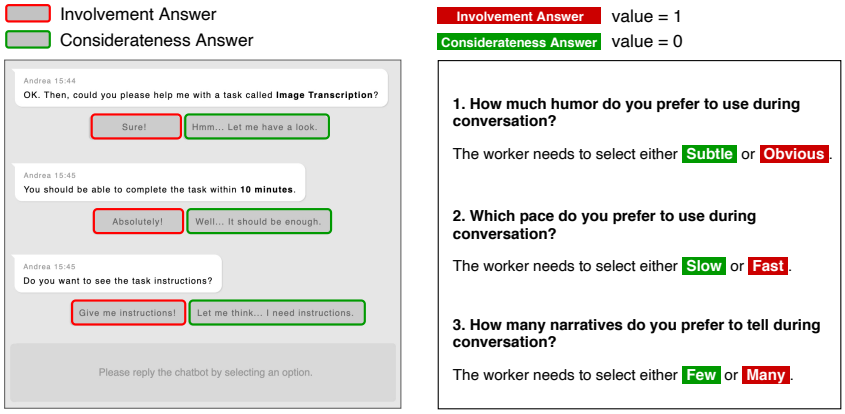
Features	**Explanation**	**Related Dimensions**
pp†	percentage of personal pronouns.	personal focus of topic
rep†	mean repeat times of words.	topic cohesion
wpu†	mean number of words per utterance.	enthusiasm
wps†	mean number of words per sentence.	enthusiasm
$nque$†	number of question marks.	use of questions
$nexcl$†	number of exclamation marks.	enthusiasm
pse†*	percentage of self-editing.	tolerance of silence
$rtype$†*	rate of typing.	pacing
$pause$†*	mean pause length.	pacing, tolerance of silence
tt†*	mean length of turn-taking pause.	pacing, tolerance of silence
$nint$†*	number of interruptions.	pacing
heu*	score from the heuristic assessment.	/
sa_h*	value of feature *humor* from self-assessment.	/
sa_p*	value of feature *pace* from self-assessment.	/
sa_n*	value of feature *narrative* from self-assessment.	/

†= features used in **Prediction with rich data**.
* = features used in **Prediction with limited data**.

In case of **Prediction with rich data**, in the preliminary conversation the conversational agent initiates a brief discussion over an engaging topic, i.e. pre-task conversation. The resulting text is labeled as either *Involvement* or *Considerateness*. The coding scheme introduced in the previous section informed the design of several textual-features (as shown in Table 2), to be automatically extracted from the pre-task conversation. These features are used for training binary classifiers.

It is neither practical nor economically viable, to ask workers to engage in relatively long pre-task conversation with the agent before each task execution - a conversational style can be contextually dependent, so it cannot be considered as an immutable property of a worker. Therefore, we investigate the effectiveness of a **Prediction with limited data** method that does not require pre-task conversation, and needs a heuristic assessment and a short self-assessment instead.

Heuristic assessment takes place during the first three interactions of a task's execution, when the conversational agent is introducing the task title, time limit and task instructions respectively, as shown in Figure 2 (a). After each interaction, the agent provides the worker with two options to select. One option corresponds to Involvement style, while the other one corresponds to Considerateness style. The result *heu* derived from the heuristic assessment is also used as a feature for style prediction, which is calculated by $heu = h1 + h2 + h3$, representing the superimposition of answers from the first, second and third interactions respectively (h_i, $1 \leq i \leq 3$, is assigned to 1 if the Involvement answer is selected by the worker, otherwise it is assigned to -1).



(a) Heuristic-assessment (b) Self-assessment

Fig. 2. (a) Heuristic-assessment and (b) Self-assessment of conversational style.

Self-assessment requires asking workers about their conversational styles. As shown in Figure 2 (b), we design a short self-assessment (that can be arranged either before or after crowdsourcing task execution as a part of pre- or post-task survey) indirectly asking workers about their preferences during the conversation, according to three continua used by Tannen in her example: *humor*, *pace* and *narrative*. Notice that these continua can be used in the scope of this self-assessment because they do not need to be inferred, but they can be directly reported by a user. The two optional answers of each question correspond to Involvement (value: 1) and Considerateness (value: 0) separately. Therefore three features pertaining to self-assessment (sa_h, sa_p and sa_n relate to humor, pace and narrative respectively) are considered in **Prediction with limited data**.

All the linguistic features for conversational style prediction (**Prediction with limited data** and **Prediction with rich data**) are listed in Table 2. Features *pp*, *rep*, *wpu*, *wps*, *nque*, and *nexcl* are only used in **Prediction with rich data**, because they are only available when the conversation

allows for subjective elements (such as personal topics, narratives, opinions and emotions) to be expressed. Features *pse*, *rtype*, *pause*, *tt* and *nint* are used in both **Prediction with rich data** and **Prediction with limited data**, since they appear in any kinds of the conversation, including the ones for microtask execution. Features *heu*, sa_h, sa_p and sa_n come from the heuristic-assessment and the self-assessment, so they are only used in **Prediction with limited data**.

The linguistic features shown in Table 2 are automatically measured and used to train supervised binary classifiers for conversational style prediction. Each training example we need in this study is a pair consisting of an a vector of linguistic features (as input) and a binary output value (1 or 0, referring to Involvement and Considerateness respectively).

5 EXPERIMENTAL SETUP

To address **RQ1**, conversational styles are independently labeled by multiple coders according to the coding scheme to understand how workers' conversational styles distribute among crowd workers, and to investigate the feasibility of style prediction using rich conversation data and limited conversation data. To address **RQ2**, we analyze the relationship between workers' conversational styles and their performance, engagement, and cognitive task load.

5.1 Conversational Task Design

The conversational crowdsourcing task (HIT) has three main phases, namely pre-task conversation, crowdsourcing microtasks, and post-task survey, as shown in Figure 3.

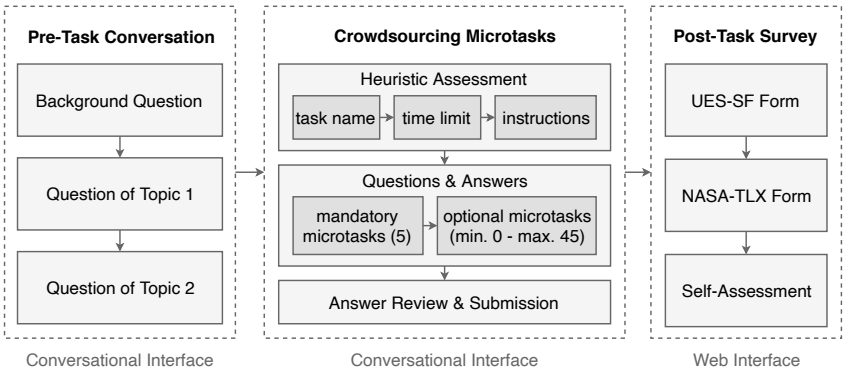


Fig. 3. Organisation of the conversational crowdsourcing task.

5.1.1 Pre-task Conversation. As mentioned in Section 4, to acquire rich data for training classifiers, the conversational agent starts a pre-task conversation encouraging workers to share about their personal stories and opinions. The conversational agent asks three questions. The first question is about the demographic background of the worker. To facilitate the production of high-quality conversation data, the second and third questions are about two controversial topics (abortion and gun control respectively).

1 *Please tell me something about yourself (your age, your gender, your ethnicity and your education background). If you want, you can also tell me about your hobbies or interesting things you would like to share.*

2 *Do you think + [**Topic 1** in Table 3]?*

3 *Please explain what you think.*

4 *Do you think + [**Topic 2** in Table 3]?*

5 *Can you explain why?*

Table 3. Controversial topics used in pre-task conversation.

Label	**Topic 1: Abortion**	**Topic 2: Gun Control**
Pro	abortion is getting rid of a fetus, not a human being reproductive choice empowers women legalizing abortion helps to reduce sexual victimization social welfare systems cannot support unwanted kids well modern abortion is safe	gun control guarantees safety of Americans guns don't kill people, people kill people free access to guns is an important right guns make up only a small part of weapons that are used to commit crimes people will always find a source for guns
Con	abortion is unsafe abortion is murder abortion damages the well-being of the mother women should not be able to use abortion as a form of contraception women should accept the consequences of pregnant	guns are an important part of the US we should control lunatics and criminals instead of guns banning guns will work not better than banning alcohol did armed guards or teachers will make schools safer gun control does not work

The rationale behind this design is that controversial topics increase interest, which also increases the likelihood of conversation [7]. Although controversial topics have also been shown to increase discomfort [7], we prevented workers from diving into a deep discussion by asking only two questions per controversial topic. Thus, controversy in general is leveraged to better stimulate the desire of expressing opinions in our scenario. The content of these two questions are picked at random from those shown in Table 3, and are inspired from recent work by Hube et al. [23]. The corresponding labels of these two questions (also shown in the table) however, are not allowed to be the same simultaneously to avoid biases. Workers are required to provide at least three sentences (each sentence contains at least two words) for each question. If the worker does not meet this requirement, the conversational agent keeps prompting with – "Can you tell me more?", "Uh huh, and?", "Good, go ahead." or other similar phrases until it receives three sentences in total.

5.1.2 Crowdsourcing Microtasks. The workflow of crowdsourcing microtasks on the conversational agent is also illustrated at the center of Figure 3.

First, the conversational agent provides workers with the basic information of the task, including task name, time limit, and task instructions. As we mentioned in the previous section (Section 4.3), the interactions at the beginning are combined with a heuristic assessment (Figure 2).

After the worker selects their preferred options, the agent proceeds to the actual task execution part – questions & answers (i.e., answering microtasks, each microtask refers to a data row or a object). The worker has to complete 5 mandatory microtasks. After the mandatory part, the worker can choose either to stay or to leave. If a worker decides to stay (i.e., continuing task execution), at most 45 optional microtasks will be presented one after another until the worker asks the agent to stop the task execution. The number of answered optional microtasks is the quantitative measurement of worker retention in our study.

After the worker decides to stop the task execution, or to complete all the optional microtasks, the conversational agent sends an answer review to the worker to check if all the previous answers are correctly recorded by the agent. Finally, after the worker has reviewed and successfully submitted the answers, the agent will send a `Task Token` to the worker. Only with this `Task Token`, the worker can proceed to complete the post-task survey.

5.1.3 Post-task Survey. User Engagement Scale Short Form [41, 42] (12 questions), NASA Task Load Index questionnaire[2] (6 questions) and the self-assessment of conversational style (3 questions) are used in the post-task survey to analyze worker engagement, cognitive task load, and conversational

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

style. First, workers have to complete the User Engagement Scale Short Form (UES-SF). The UES-SF consists of 12 questions in four factors (Focused Attention, Perceived Usability, Aesthetic Appeal, and Reward Factor). Workers are asked to answer each question by setting a slider on a 7-point sliding bar ranging from "*1: Strongly Disagree*" to "*7: Strongly Agree*".

Then, workers are asked to complete the NASA Task Load Index (NASA-TLX) questionnaire, which contains six items spanning six aspects (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration respectively). Workers need to indicate their perceived task loads on these six aspects by setting the slider to on a 20-point sliding bar ranging from "*1: Very Low*" to "*21: Very High*". The TLX scores will be scaled to 0-100 in the evaluation.

After UES-SF and NASA-TLX, workers are asked to complete the self-assessment of conversational style that has been described in section 4.3.

5.2 Independent and Dependent Variables

5.2.1 Independent Variables. Considering crowdsourcing tasks have different levels of difficulty and complexity [58], we model task difficulty into 3 levels, from easy to difficult, to observe how crowd workers with different conversational styles perform on different difficulty levels.

We consider two data types (text and image) of microtask, Information Finding and Image Transcription. We used these task types since they are representative of typical classes of microtasks [15], and they easily allow the classification of the task objective into different difficulty levels. This results in six experimental conditions (2×3), i.e. 2 types of task (Information Finding and Image Transcription, representing text-based and image-based tasks respectively) with three levels of difficulty (Level 0, Level 1 and Level 2, from easy to difficult).

1) Information Finding Tasks require workers to find the middle name of a famous person from either Wikipedia[3] or Google[4] [14]. As shown in Figure 4, this type of tasks has three different difficulty levels. In **Level 0**, the conversational agent shows first and last names of a unique, famous, and unambiguous person, whose middle names can be easily found from a search engine. In **Level 1**, the agent additionally shows the profession of the given person based on Level 0. We manually selected the names in this difficulty level to ensure there are at least two different famous persons with the same first and last names, so that the worker needs to distinguish the given person according the profession. In **Level 2**, the agent also shows the famous person's active year, based on Level 1 (showing first/last names and profession). In this difficulty level, there are multiple famous people with the same first/last name, and with the same profession. The worker is asked to find the correct one from those famous people by making use of all the given information.

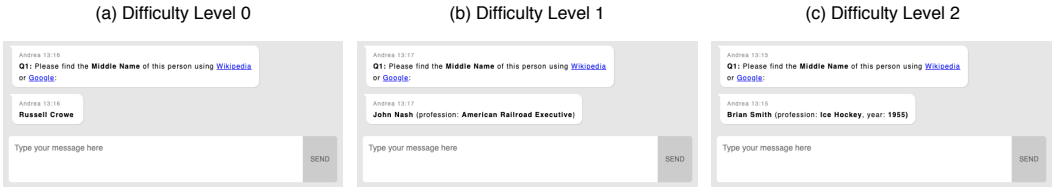


Fig. 4. Examples of the Information Finding task with three difficulty levels.

2) Image Transcription Tasks require workers to read the image (randomly generated by Claptcha[5]) and transcribe it into letters, as shown in Figure 5. This task type also has three different levels of the

[3] http://en.wikipedia.org/
[4] http://www.google.com/
[5] https://github.com/kuszaj/claptcha

difficulty. In **Level 0**, the conversational agent shows an image containing a random combination of 6 letters (from the English alphabet, either in the upper or lower cases) with the highest resolution (approximately 38 px × 75 px per letter). The least noises (Gaussian white noises, $\sigma = 0.1$) are added into the image. In **Level 1**, the agent shows an image containing a random combination of 12 random letters with the medium resolution (35 px × 70 px per letter). The medium noises (Gaussian white noises, $\sigma = 0.3$) are added into the image. In **Level 2**, the agent shows an image containing a random combination of 18 random letters with the lowest resolution (approximately 33 px × 65 px px per letter). The largest noises (Gaussian white noises, $\sigma = 0.5$) are added into the image.

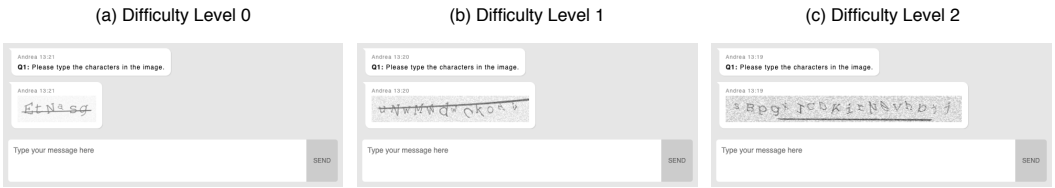


Fig. 5. Examples of the Image Transcription task with different difficulty levels.

5.2.2 Dependent Variables. We measure the performance of conversational style prediction, output quality, worker engagement and cognitive task load. We use these metrics to analyse what role the conversational style plays in microtask crowdsourcing.

1) Performance of conversational style prediction is measured by comparing prediction results of classifiers with the labeled ground truth produced based on the coding scheme (the coding process is explained in section 6.1). We measure the overall accuracy, noted as *accuracy*, which is the fraction of correctly predicted conversation style of workers among all the workers.

2) Output quality is measured using worker accuracy, which is calculated by comparing the answers provided by workers with the expected value:

$$\text{output quality} = \frac{\text{\# correct answers provided by the worker}}{\text{\# all the provided answers}} \tag{1}$$

In case of Information Finding tasks, the answer provided by workers should contain the expected value (case insensitive). In terms of Image Transcription tasks, the *string similarity* between the answer provided by workers and the expected value should be greater than 0.8. The string *similarity* is calculated as $similarity = 2 \times M/T$, where T is the total number of characters in both answers and ground truth, and M is the number of matched characters. Note that the *similarity* equals to 1 if the answer is identical to the ground truth (case insensitive), and equals to 0 if there is nothing in common (case insensitive).

3) Worker engagement is measured using two approaches: the first one is worker *retention*, quantified by the number of optional microtasks answered (from 0 to 45); and the second is the short-form of the user-engagement scale [42] – *UES-SF* scores in four different factors (Focused Attention, Perceived Usability, Aesthetic Appeal, and Reward Factor), ranging from 1 to 7. A relatively higher UES-SF score indicates that the worker is more engaged with regard to the corresponding factor [62].

4) Cognitive task load of workers is evaluated by using the 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. Higher the TLX score is, the heavier task load the worker perceives.

5.3 Experimental Environment

5.3.1 Workers. There are 6 experimental conditions (2 task types $\times$ 3 difficulty levels), and each experimental condition has 50 crowdsourcing microtasks. As each microtask requires answers from at least 3 unique workers and each worker must submit at least 5 mandatory microtasks, we recruited $50 \times 3/5 = 30$ unique workers for each experimental condition from Amazon Mechanical Turk, resulting in $30 \times 6 = 180$ unique workers for the entire experiment. Only crowd workers whose HIT approval rates are greater than 95% could view and accept our crowdsourcing tasks [11].

5.3.2 Rewards. To avoid monetary biases, we immediately pay 1 USD to each worker after the worker submits the task (HIT). Then all the workers equally receive 0.01 USD for each optional Image Transcription microtask, or receive 0.02 USD for each optional Information Finding microtask according to how many optional microtasks they answer after we approve their submissions. To ensure that we consistently pay an average hourly wage of 7.5 USD, we additionally bonus workers after calculating their execution time.

5.3.3 Quality Control. To avoid biases caused by repeated task execution [14], we added extra Javascript code for recording their worker IDs on our server, to prevent workers from executing HITs multiple times. If a worker attempts to complete another HIT in this batch after having one already (meaning his/her worker ID has been recorded), all the instructions and questions on the web page are removed. Instead, a message, that kindly informs workers that they should "return" because of our experimental restrictions, is displayed on the task page.

6 EVALUATION

6.1 Conversational Style Estimation

Coding conversational styles. With this evaluation we address **RQ1**. The coding process is conducted by three of our authors, who have deeply studied the theory of conversational style and understood the concept of linguistic devices. The inter-rater reliability is measured by Fleiss' Kappa. Three coders are in complete agreement for 124 out of 180 crowd workers. The 56 cases having disagreement are disambiguated manually by coders. In total, 86 workers exhibited *Involved* style, while 94 workers showed *Considerate* style. Therefore the kappa κ value is 0.78.

The score distributions (of three coders, 180 judgments per dimension) of five dimensions are shown in Figure 6. *Pacing* and *Tolerance* are automatically calculated according to Table 1, therefore scores (-1 and +1) of these two dimensions are equally distributed. *Personal focus of topic*, *Enthusiasm* and *Topic cohesion & use of questions* are manually labeled by coders. As we can see from the Figure, Coder1 tended to give lower scores (more considerate) while Coder3 tended to give higher scores (more involved). However, scores given by different coders are distributed similarly in general (with only small fluctuations), suggesting that the coding scheme is sufficiently robust to account for the subjectivity of the coders.

Prediction accuracy. The training dataset (features extracted by the conversational agent with ground truth data labeled by the coding scheme) was randomly divided into two part - 70% of them went to the training set, while the rest (30%) went to the testing set.

Because the focus of this paper is the feasibility of the conversational style prediction, we did not use the state-of-the-art classifiers (e.g. deep neural network) and attempt to improve their performance. Instead, we only select most basic and naive classifiers (Logistic Regression, Naive

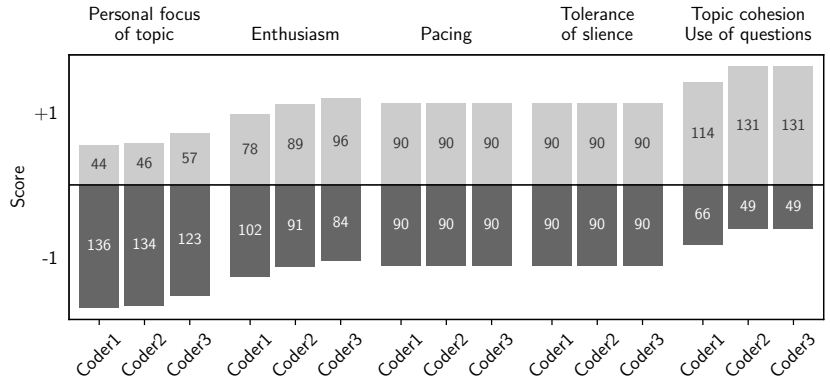


Fig. 6. The score distribution of three coders across five dimensions. Numbers in this figure represent the number of judgments given by the corresponding coder.

Bayes, k-Nearest Neighbors, Decision Tree, Gradient Boosting), and compare the accuracy[6] to understand the feasibility of predicting conversational styles.

The prediction performance is shown in Table 4. Gradient Boosting can obtain highest performance by using **Prediction with rich data**, whose overall accuracy value is nearly 80%. These results are encouraging, showing that it is indeed possible to achieve good prediction performance for conversational style using rich conversational data. In terms of **Prediction with limited data**, once again, Gradient Boosting can obtain highest performance, whose overall accuracy reaches 66.7%. These results are also encouraging, as they suggest that conversational styles can be predicted also through limited interactions with the workers. Since we didn't study deep into the parameter adjustment of machine learning models, we believe a well-tuned prediction method with limited data still has great potential to accurately estimate the conversational style.

Table 4. The accuracy (unit: percent %) of conversational style prediction by common classifiers, where the classifier with the highest accuracy is highlighted.

Prediction methods	**Logistic Regression**	**Naive Bayes**	**k-NN**	**Decision Tree**	**Gradient Boosting**
with rich data	75.93	75.93	74.07	77.78	79.63
with limited data	57.41	55.56	48.15	59.26	66.67

6.2 The Performance of Workers with Two Conversational Styles

Execution time. *Workers with an Involvement style generally feature less task execution time.* The average execution time of pre-task conversation, heuristic assessment & self-assessment, crowdsourcing microtasks, and USE-SF & NASA-TLX surveys are shown in Table 5. As workers with an Involvement style have a faster pace by definition, their task execution time are generally less than Considerate workers. We highlight that the average time spent on heuristic assessment and self-assessment is 73 seconds (around 0.15 USD per worker), while the average time spent on pre-task conversation is 412 seconds (around 0.86 USD per worker), where the latter is 5.6

[6]Details of the classifiers and other prediction metrics such as precision, recall and F1 are explained in the supplementary material.

times longer than the former, meaning Prediction with rich data is 5.6 times more expensive than Prediction with limited data.

Table 5. The execution time ($\mu \pm \sigma$, unit: second) of pre-task conversation, heuristic assessment & self-assessment, crowdsourcing microtasks, and USE-SF & NASA-TLX surveys.

Conversational styles	Pre-task conversation	Heuristic- & self-assessment	Crowdsourcing microtasks	USE-SF & NASA-TLX	Total
Involvement	376 ± 270	69 ± 55	266 ± 231	129 ± 179	842 ± 459
Considerateness	444 ± 247	77 ± 56	318 ± 297	153 ± 251	994 ± 551
Overall	412 ± 260	73 ± 56	293 ± 269	142 ± 220	922 ± 515

Output quality. *Workers with an Involvement style have significantly higher output quality at high difficulty level compared to Considerate workers.* We calculate the output quality (worker accuracy) across all types of tasks and difficulty levels, which are listed in Table 6. Obviously, the overall worker accuracy decreases as the task difficulty increases across two task types. We observe that when the overall worker accuracy is lower than 60% (IF Level 1, IF Level 2 and IT Level 2), the workers with *Involvement* style have higher average accuracy than the workers of *Considerateness* style. As the worker accuracy does not follow the normal distribution according to Shapiro-Wilk tests ($\alpha = 0.05$), with the Wilcoxon Rank-Sum test ($\alpha = 0.05$) we tested the significance of pairwise differences between two conversational styles. We found statistical significance for all the Level 2 Tasks (both Information Finding and Image Transcription, $p = 0.045$ and $p = 0.021$ respectively).

Table 6. The worker accuracy ($\mu \pm \sigma$: mean and standard deviation, unit: percentage) of all task types with difficulty levels.

Task type	Difficulty	Involvement	Considerateness	Overall
Information Finding	Level 0	66.76 ± 38.84	73.77 ± 33.17	69.8 ± 36.66
	Level 1	54.05 ± 35.1	50.42 ± 35.22	52.11 ± 35.21
	Level 2*	55.0 ± 33.84	25.95 ± 28.84	36.3 ± 33.0
Image Transcription	Level 0	86.48 ± 22.91	90.56 ± 8.31	87.7 ± 19.79
	Level 1	76.92 ± 29.19	79.41 ± 26.0	78.33 ± 27.46
	Level 2*	63.18 ± 18.86	35.68 ± 32.74	45.77 ± 31.39

* = statistically significant (Involvement vs Considerateness).

Worker engagement: Worker retention. *We found no significant difference in worker retention.* We counted optional microtasks that workers answered, and plotted distributions of number of answered optional microtasks across different task types and difficulty levels using a violin plot (Figure 7). In this figure, the width of each violin at any point represents the number of workers that answered the corresponding number of optional microtasks. According to the shape of each "violin", the distributions do not meet any assumptions for parametric tests (also verified by Shapiro-Wilk tests), therefore we use the Wilcoxon Rank-Sum test (two-tailed, $\alpha = 0.05$) to test the significance of difference between two conversational styles. We found no significant difference in worker retention (answered optional microtasks) between the workers with two conversational styles.

Worker engagement: UES-SF score. *Workers with an Involvement style reported significantly higher scores on UES-SF questionnaire in most cases of specific UES factors compared to Considerate workers.*

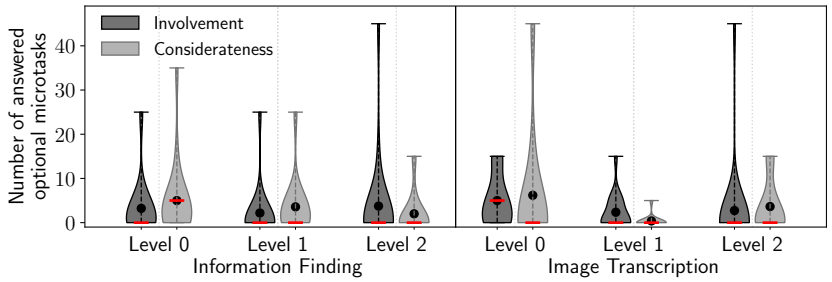


Fig. 7. A violinplot representing the number of optional microtasks answered by workers across different task types and difficulty levels, where the red lines represent the median value and the black points represent the mean value.

UES-SF scores of four factors as well as overall scores are plotted in Figure 8. Since the distributions of UES-SF scores meet the assumption of normal distributions according to Shapiro-Wilk tests, to study user engagement of workers with different conversational styles, t-tests (two tailed, $\alpha = 0.05$) were performed to test the significance of differences between two conversational styles. Because multiple comparisons (different UES factors) are conducted between two datasets (two conversational styles), Bonferroni correction is used to control Type-I error.

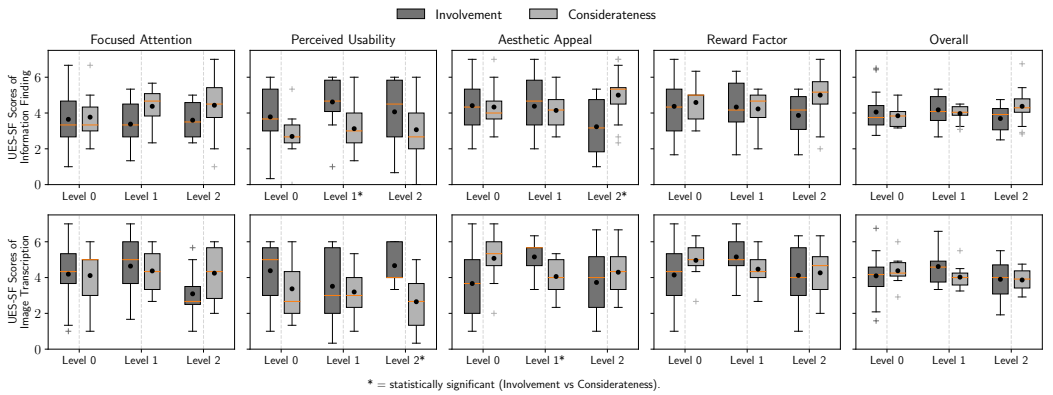


Fig. 8. Boxplots of UES-SF scores by Task-Difficulty and Conversational Style, where the red lines represent the median value and the black points represent the mean value.

In terms of Information Finding tasks, whose UES-SF scores are displayed in the first row of Figure 8, workers of Involvement style reported significantly higher scores with respect to Perceived Usability factor, when difficulty level is 1 (IF Level 1 on Perceived Usability, $p = 0.0033$). Workers of Considerateness style reported higher Aesthetic Appeal score at level 2 (IF Level 2 on Aesthetic Appeal, $p = 0.0026$). As for Image Transcription tasks (UES-SF scores are displayed in the second row of Figure 8), workers of Involvement style gave higher scores on Perceived Usability and Aesthetic Appeal when difficulty level is 2 and 1 respectively (IT Level 2 on Perceived Usability and IT Level 1 on Aesthetic Appeal, $p = 0.0003$ and $p = 0.0034$ respectively). We found no significant difference between two styles in terms of Overall UES-SF score.

To conclude, workers with an Involvement style tended to report higher scores on UES-SF questionnaire, while workers with both styles did not show significant differences with respect to worker retention in both tasks.

Cognitive task load. *Workers with an Involvement style reported significantly less cognitive task load at high difficulty level compared to Considerate workers.* Results of unweighted NASA-TLX scores are displayed as box plots in Figure 9. As we can see from the figure, workers of Involvement style reported lower mean values than workers of Considerateness style in all the experimental conditions except only one case (IT Level 1). Similarly, those Involvement workers also reported lower median values across all the experimental conditions except the same case (IT Level 1).

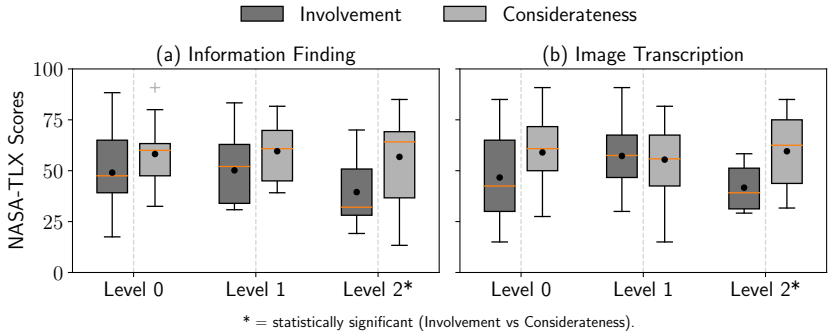


Fig. 9. Boxplots of NASA-TLX scores by Task-Difficulty and Conversational Style, where the red lines represent the median value and the black points represent the mean value.

Since the distributions of TLX scores meet the assumption of normal distributions according to Shapiro-Wilk tests, we conducted t-tests (two tailed, $\alpha = 0.05$) to find significant pairwise differences between two conversational styles across different task types and difficulty levels. Results show that workers of Involvement style reported less cognitive task load than workers of Considerateness style, after they completed all the tasks of Level 2 with significant differences (both Information Finding and Image Transcription, $p = 0.034$ and $p = 0.006$ respectively). These results collectively suggest that workers of Involvement style perceive less task load from task execution than workers of Considerateness style in general, especially when completing difficult tasks.

7 DISCUSSION

7.1 Reflection on Conversational Style Estimation

In this paper, we proposed a coding scheme for conversational style estimation that systematically classifies the text-based conversational style into two categories - *Involvement* and *Considerateness*. To the best of our knowledge, this is the first work that proposes a systematic method to estimate the conversational style of online workers. We also studied the feasibility of automatically predicting workers' conversational styles by common classifiers using rich conversational data (more accurate) and limited conversational data (less expensive) respectively. Results show that indeed conversational styles can be predicted, using the proposed task design, thus showing that conversational style estimation can serve as a practical tool for microtask crowdsourcing.

7.2 Reflection on the Performance of Workers

We explored the behaviour of online workers with two conversational styles during conversational microtasking, and observed strong evidence that conversational style could bear relationship with

quality of outcome for difficult tasks. We found that Involvement workers performed better in terms of quality-related outcomes in tasks with higher difficulty levels. These results suggest that conversational style estimation could be a useful tool for output quality prediction. Analysis of cognitive task load revealed that workers of Involvement style perceived less task load with higher difficulty levels. Our experimental findings suggest that the conversational style estimation can be used for worker performance prediction, to better enable adaptive crowdsourcing strategies.

7.3 Implications for Design

The results show that the conversational style has a significant impact on output quality, worker engagement, and cognitive task load. Workers with an Involvement style can produce higher work accuracy, perceive higher user engagement and feel less cognitive task load when they are completing difficult tasks. This gives us a strong indication that conversational style estimation could be an effective tool for predicting worker performance and assisting crowdsourcing task assignment. Highly involved workers could be selected and assigned to tasks of higher complexity, to produce high-quality work outcomes. The proposed method can be applied in microtask crowdsourcing marketplaces to improve work outcomes and foster a better worker-requester relationship (by improving worker accuracy, increase worker engagement, and reducing cognitive task load).

We found that a long conversation (pre-task conversation) can provide enough data for precisely (80%) predict the conversational style, however, it takes 5.6 times longer than a heuristic assessment with a self-assessment (67% and only takes around 1 minute). The precision of prediction with limited data is lower but acceptable, and we are confident that it could be improved using state-of-art classifiers. We suggest that in the future design of the conversational microtask crowdsourcing, a heuristic assessment with a self-assessment could be considered as an extra test to effectively predict worker performance (which also help in dynamically adapting task scheduling and assignment strategies), while it only costs 0.1 to 0.2 USD additionally on each worker.

This work propose a conversational style estimation method for conversational microtask crowdsourcing. Since previous studies have shown the feasibility of deploying microtasks on Facebook (Messenger) [5, 18, 54], Twitter [46], etc., the proposed method can be applied in chatbot systems on common social network channels for analyzing and understanding the personality, mood, subjectivity, and bias of (particularly anonymous) users.

7.4 Caveats, Limitations and Future Work

In terms of style estimation and prediction, we only focus on the text-based conversation. Text-based conversation ignores several paralinguistic features (pitch, voice) and nonlinguistic features (smile, laughter, gestures). Moreover, some features relying on the analysis of voice such as pacing were measured in a different way. In general, there are various means to interact with conversational agents (e.g., voice-based agent, video-based agent). Conversational agents and corresponding style estimation methods based on voice or video could be an interesting direction to explore. Second, there is still room for improvement for style prediction. In this study, we used the most basic and naive binary classifiers for style prediction, and did not tune the parameters of those classifiers to pursue higher accuracy. Thus, in the imminent future, state-of-the-art machine learning methods can be applied for conversation style prediction. The process of labeling ground truth data by using the coding scheme can also be crowdsourced as HITs, to label larger training datasets with the cooperative work of crowd workers.

In terms of the study of the impact of conversational styles, the task types studied in the experiment are limited. We only focus on one input type of microtask – that is free text. In the case of this input type, workers are asked to provide their responses using free text (string). However, many other input types such as multiple choices, sliding bars, and bounding boxes are also used

in microtasks of current online crowdsourcing marketplaces. Studying the performance of crowd workers with different conversational styles on other types of tasks is an important next step to our work. Previous work shows that an aligned style of the conversational agent can improve worker performance [48, 52]. Future experiments should consider assigning different conversational styles to the conversational agent, and investigate whether an aligned style can help in the improvement of worker engagement and reduction of cognitive task load.

Moreover, to estimate conversational styles, the crowd workers in our experiments spent a long time on pre-task conversations (around 7 minutes), which might have a negative impact on work outcomes. A future study can explore style prediction and alignment on the experimental conditions without additional conversation to further evaluate the breadth of practical value in conversational style prediction.

The estimation and prediction of the conversational style of a worker has clear implications for privacy and fairness. While performing a dialogue, workers could disclose personal information that should be treated according to best practices and regulation for personal data management. In this respect, the *Prediction with limited data* allows for a data collection method that is minimally invasive and that could be easily repeated across HITs, thus eliminating the need for storage and management of worker data while achieving good prediction performance. We also stress the potential fairness implications that conversational style profiling can have on task distributions and therefore, on the earning power of workers. We believe that user modelling approaches like the one we propose should be used only in the context of an explicit, optional, and informed approval from workers, and such that a lack of approval does not lead to overall compensation penalties.

Future work can focus on the usage of a variety of different types of conversational agents, the collection of large amount of conversational data for constructing a training dataset and deep-learning-based classifiers for conversational style estimation. Various crowdsourcing task types and conversational style aligning strategies could be considered in the future.

8 CONCLUSIONS

In this work, we explored how the conversational style of a crowd worker could be reliably estimated during crowdsourcing microtasking, and the relationship between conversational styles and quality-related outcomes, worker engagement and cognitive task load. We addressed **RQ1** by proposing a coding scheme for style estimation based on the five dimensions of examined linguistic devices, and style prediction methods using either rich conversational data or limited conversational data. We answered **RQ2** by performing a crowdsourcing experiment to analyze the behaviour of crowd workers with different conversational styles during crowdsourcing microtasking in terms of the output quality, worker engagement and cognitive task load.

We found that common classifiers could obtain a reasonably good accuracy in conversation style prediction. Furthermore, experimental results revealed that workers' Involvement conversational style relates with higher output quality, higher user engagement and less perceived task load in tasks with higher difficulty.

REFERENCES

[1] Ahmed Abbasi, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)* 26, 3 (2008), 1–34.

[2] Sandeep Avula, Gordon Chadwick, Jaime Arguello, and Robert Capra. 2018. SearchBots: User Engagement with ChatBots During Collaborative Search. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. ACM, 52–61.

[3] Alessandro Bozzon, Piero Fraternali, Luca Galli, and Roula Karam. 2014. Modeling CrowdSourcing Scenarios in Socially-Enabled Human Computation Applications. *Journal on Data Semantics* 3, 3 (2014), 169–188. https://doi.org/10.1007/s13740-013-0032-2

[4] Luka Bradeško, Michael Witbrock, Janez Starc, Zala Herga, Marko Grobelnik, and Dunja Mladenić. 2017. Curious Cat–Mobile, Context-Aware Conversational Crowdsourcing Knowledge Acquisition. *ACM Transactions on Information Systems (TOIS)* 35, 4 (2017), 33.

[5] Erin Brady, Meredith Ringel Morris, and Jeffrey P Bigham. 2015. Gauging receptiveness to social microvolunteering. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1055–1064.

[6] John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 1301–1309.

[7] Zoey Chen and Jonah Berger. 2013. When, why, and how controversy causes conversation. *Journal of Consumer Research* 40, 3 (2013), 580–593.

[8] Phil Cohen, Adam Cheyer, Eric Horvitz, Rana El Kaliouby, and Steve Whittaker. 2016. On the future of personal assistants. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1032–1037.

[9] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. What can i help you with?: infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 43.

[10] Gianluca Demartini, Djellel Eddine Difallah, Ujwal Gadiraju, Michele Catasta, et al. 2017. An introduction to hybrid human-machine information systems. *Foundations and Trends® in Web Science* 7, 1 (2017), 1–87.

[11] Carsten Eickhoff and Arjen P de Vries. 2013. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval* 16, 2 (2013), 121–137.

[12] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[13] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. 2017. Modus operandi of crowd workers: The invisible role of microtask work environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 49.

[14] Ujwal Gadiraju and Stefan Dietze. 2017. Improving learning through achievement priming in crowdsourced information finding microtasks. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. 105–114.

[15] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. 2014. A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM, 218–223.

[16] Amy L Gonzales, Jeffrey T Hancock, and James W Pennebaker. 2010. Language style matching as a predictor of social dynamics in small groups. *Communication Research* 37, 1 (2010), 3–19.

[17] Bettina Graf, Maike Krüger, Felix Müller, Alexander Ruhland, and Andrea Zech. 2015. Nombot: simplify food tracking. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia*. ACM, 360–363.

[18] Nathan Hahn, Shamsi T Iqbal, and Jaime Teevan. 2019. Casual Microtasking: Embedding Microtasks in Facebook. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 19.

[19] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. All those wasted hours: On task abandonment in crowdsourcing. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 321–329.

[20] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. The impact of task abandonment in crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering* (2019).

[21] Ting-Hao Kenneth Huang, Joseph Chee Chang, and Jeffrey P Bigham. 2018. 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*. ACM, 295.

[22] Ting-Hao Kenneth Huang, Walter S Lasecki, and Jeffrey P Bigham. 2015. Guardian: A crowd-powered spoken dialog system for web apis. In *Third AAAI conference on human computation and crowdsourcing*.

[23] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[24] Kevin Jepson. 2005. ConversationsâĂŤand negotiated interactionâĂŤin text and voice chat rooms. *Language Learning & Technology* 9, 3 (2005), 79–98.

[25] Patrik Jonell, Mattias Bystedt, Fethiye Irmak Dogan, Per Fallgren, Jonas Ivarsson, Marketa Slukova, José Lopes Ulme Wennberg, Johan Boye, and Gabriel Skantze. 2018. Fantom: A Crowdsourced Social Chatbot using an Evolving Dialog Graph. *Proc. Alexa Prize* (2018).

[26] Joy Kim, Sarah Sterman, Allegra Argent Beal Cohen, and Michael S Bernstein. 2017. Mechanical novel: Crowdsourcing complex work through reflection and revision. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 233–245.

[27] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 86, 12 pages. https://doi.org/10.1145/3290605.3300316

[28] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. ACM, 121–130.

[29] Aniket Kittur, Susheel Khamkar, Paul André, and Robert Kraut. 2012. CrowdWeaver: visually managing complex crowd work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 1033–1036.

[30] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 43–52.

[31] Ari Kobren, Chun How Tan, Panagiotis Ipeirotis, and Evgeniy Gabrilovich. 2015. Getting More for Less: Optimized Crowdsourcing with Dynamic Tasks and Goals. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 592–602. https://doi.org/10.1145/2736277.2741681

[32] Pavel Kucherbaev, Alessandro Bozzon, and Geert-Jan Houben. 2018. Human-Aided Bots. *IEEE Internet Computing* 22, 6 (2018), 36–43.

[33] Robin Tolmach Lakoff. 1979. Stylistic strategies within a grammar of style. *Annals of the New York Academy of Sciences* 327, 1 (1979), 53–78.

[34] Walter S Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F Allen, and Jeffrey P Bigham. 2013. Chorus: a crowd-powered conversational assistant. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. ACM, 151–162.

[35] Ewa Luger and Abigail Sellen. 2016. Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5286–5297.

[36] Andrew Mao, Ece Kamar, and Eric Horvitz. 2013. Why stop now? predicting worker engagement in online crowdsourcing. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[37] Panagiotis Mavridis, Owen Huang, Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2019. Chatterbox: Conversational Interfaces for Microtask Crowdsourcing. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, 243–251.

[38] Robert J Moore, Raphael Arar, Guang-Jie Ren, and Margaret H Szymanski. 2017. Conversational UX design. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 492–497.

[39] Sean A Munson, Karina Kervin, and Lionel P Robert Jr. 2014. Monitoring email to indicate project team performance and mutual attraction. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 542–549.

[40] Dong Nguyen, Noah A Smith, and Carolyn P Rosé. 2011. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*. Association for Computational Linguistics, 115–123.

[41] Heather O'Brien. 2016. Theoretical perspectives on user engagement. In *Why Engagement Matters*. Springer, 1–26.

[42] Heather L O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28–39.

[43] Nilma Perera, Gregor Kennedy, and Jon Pearce. 2008. Are You Bored?: Maybe an Interface Agent Can Help!. In *Proceedings of the 20th Australasian Conference on Computer-Human Interaction: Designing for Habitus and Habitat (OZCHI '08)*. ACM, New York, NY, USA, 49–56. https://doi.org/10.1145/1517744.1517760

[44] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Improving Worker Engagement Through Conversational Microtask Crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–12.

[45] M Rafael Salaberry. 2000. L2 morphosyntactic development in text-based computer-mediated communication. *Computer Assisted Language Learning* 13, 1 (2000), 5–27.

[46] Saiph Savage, Andres Monroy-Hernandez, and Tobias Höllerer. 2016. Botivist: Calling volunteers to action using online bots. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 813–822.

[47] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging.. In *AAAI spring symposium: Computational approaches to analyzing weblogs*. 199–205.

[48] Ameneh Shamekhi, Mary Czerwinski, Gloria Mark, Margeigh Novotny, and Gregory A Bennett. 2016. An exploratory study toward the preferred conversational style for compatible virtual agents. In *International Conference on Intelligent*

Virtual Agents. Springer, 40–50.
[49] Deborah Tannen. 1987. Conversational style. *Psycholinguistic models of production* (1987), 251–267.
[50] Deborah Tannen. 2005. *Conversational style: Analyzing talk among friends*. Oxford University Press.
[51] Yla R Tausczik and James W Pennebaker. 2013. Improving teamwork using real-time language feedback. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 459–468.
[52] Paul Thomas, Mary Czerwinski, Daniel McDuff, Nick Craswell, and Gloria Mark. 2018. Style and alignment in information-seeking conversation. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. ACM, 42–51.
[53] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2017. MISC: A data set of information-seeking conversations. In *SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*, Vol. 5.
[54] Carlos Toxtli, Joel Chan, Walter S Lasecki, and Saiph Savage. 2018. Enabling Expert Critique with Chatbots and Micro Guidance. In *Collective Intelligence 2018*. ACM, 4.
[55] Melissa A Valentine, Daniela Retelny, Alexandra To, Negar Rahmati, Tulsee Doshi, and Michael S Bernstein. 2017. Flash organizations: Crowdsourcing complex work by structuring crowds as organizations. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. ACM, 3523–3537.
[56] Bert Vandenberghe. 2017. Bot personas as off-the-shelf users. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 782–789.
[57] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles LA Clarke. 2017. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2187–2193.
[58] Jie Yang, Judith Redi, Gianluca Demartini, and Alessandro Bozzon. 2016. Modeling task complexity in crowdsourcing. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.
[59] Xi Yang, Marco Aurisicchio, and Weston Baxter. 2019. Understanding Affective Experiences With Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 542.
[60] Ying Zhang, Xianghua Ding, and Ning Gu. 2018. Understanding Fatigue and its Impact in Crowdsourcing. In *2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD))*. IEEE, 57–62.
[61] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. 2019. Dissonance Between Human and Machine Understanding. *Proc. ACM Hum.-Comput. Interact. 3 CSCW* 56 (2019), 26.
[62] Mengdie Zhuang and Ujwal Gadiraju. 2019. In What Mood Are You Today? An Analysis of Crowd Workers' Mood, Performance and Engagement. In *Proceedings of the 10th ACM Conference on Web Science*. 373–382.

Received October 2019; revised January 2020; accepted March 2020