Automated Writing Assistance for Task Description Clarity in Crowdsourcing

From the
Faculty of Computer Science, Electrical Engineering, and Mathematics
of Paderborn University

The submitted dissertation of
Zahra Nouri
to obtain the academic degree of
Dr. rer. nat.

Paderborn, Germany
March 2024

Dissertation

Automated Writing Assistance for Task Description Clarity in Crowdsourcing

Zahra Nouri, Paderborn University

Paderborn, Germany, 2024

Reviewers

Prof. Dr. Henning Wachsmuth, Leibniz University Hannover

Prof. Dr. Ir. Ujwal Gadiraju, Delft University of Technology

Doctoral Committee

Prof. Dr. Henning Wachsmuth, Leibniz University Hannover

Prof. Dr. Ir. Ujwal Gadiraju, Delft University of Technology

Prof. Dr. Gregor Engels, Paderborn University

Prof. Dr. Johannes Blömer, Paderborn University

Prof. Dr. Eckhard Steffen, Paderborn University

To my dearest family,

This thesis is lovingly dedicated to my mother, who has been the bedrock of my life. Her unwavering belief in me and her invaluable moral support have given me the courage to reach for the stars. She has taught me how to be resilient and persevere, even when the odds seemed insurmountable.

Thank you, Mom, for instilling in me the virtues and values that have shaped who I am today.

To my dear sister and brothers, my closest allies and partners in crime. Without their guidance and encouragement, this journey would have been a much more difficult one. Thank you for always being there for me, through thick and thin.

The achievements within these pages are not solely mine, but also belong to them. Their love and support have nurtured my growth, both personally and academically, and I am forever grateful.

Contents

Preface

Abstract

The quality of solutions submitted by workers on crowdsourcing platforms is influenced by problems that they encounter during their work in crowdsourcing marketplace. In this thesis, our initial objective is to identify the primary obstacle that has the most significant influence on maximizing the benefits derived from the crowdsourcing model. Subsequently, we will employ an approach to investigate whether the use of natural language processing techniques can enhance the identified issue. To find the primary problems, we offer a brief yet comprehensive survey based on two complementary investigations: (1) a *literature review*, in which we arrange problems identified through interviews with workers, and (2) an empirical *data analysis*, where topic modeling is applied to extract workers' grievances from an English corpus of workers' forum discussions. The literature discusses task evaluation issues as most widespread, while the data suggests that poor task design by requesters appears to be the most troubling issue for workers in the crowdsourcing processes. Prior research shows that inexperienced requesters fail to write clear and complete task descriptions which directly leads to low quality submissions from workers. In this thesis, we aim to address this issue by studying whether an automated writing assistance can enable requesters to identify and improve clarity flaws in their task descriptions before deployment on the platform. In order to achieve this, we undertake two significant measures: a) we first study whether clarity flaws in task descriptions can be identified automatically using natural language processing methods. We identify and synthesize eight clarity flaws in task descriptions and we then build both BERT-based and feature-based binary classifiers. Through a crowdsourced study, we collect annotations of clarity flaws in 1332 real task descriptions. Using this dataset, we evaluate several configurations of the classifiers. Our results indicate that nearly all the clarity flaws in task descriptions can be assessed reasonably well by the classifiers. b) Upon the insights achieved from previous step, we developed a tool that enables requesters to iteratively identify and correct eight common clarity flaws in their task descriptions before deployment on platforms. It employs natural language processing models trained on real-world task descriptions that score a given task description for the eight clarity flaws. In a two-phase user study, we evaluate whether automated assis-

tance for writing task descriptions proves beneficial from the requesters' viewpoint and effective from the perspective of crowd workers, as those who are confronted with such descriptions in practice. Based on our findings, approximately 65% of requesters rated the tool's information assistance as highly or very highly helpful. Furthermore, 76% of crowd workers reported an improvement in the overall clarity of task descriptions when requesters utilized the tool. The results indicate that by employing natural language processing techniques, we can automatically aid requesters in identifying clarity issues within their task descriptions, leading to enhancements that make them clearer for workers. This, in turn, results in improved task design quality and subsequently, addresses workers' submissions quality as a major challenge in crowdsourcing processes.

Zusammenfassung

Die Qualität der Lösungen, die von Arbeitnehmern auf Crowdsourcing-Plattformen eingereicht werden, wird von den Herausforderungen beeinflusst, mit denen sie während ihrer Arbeit auf dem Crowdsourcing-Markt konfrontiert werden. In der vorliegenden Dissertation besteht unser erstes Ziel darin, das Hauptproblem zu identifizieren, das den größten Einfluss auf die Maximierung der Vorteile des Crowdsourcing-Modells hat. Anschließend werden wir einen Ansatz verwenden, um zu untersuchen, ob der Einsatz von Techniken zur natürlichen Sprachverarbeitung das identifizierte Problem verbessern kann. Um die Hauptprobleme zu finden, bieten wir eine kurze, aber umfassende Umfrage auf der Grundlage von zwei ergänzenden Untersuchungen an: (1) eine Literaturübersicht, in der wir die Herausforderungen identifizieren, die durch Interviews mit Arbeitnehmern ermittelt wurden, und (2) eine empirische Datenanalyse, bei der Topic Modeling angewendet wird, um Beschwerden von Arbeitnehmern aus einem englischen Korpus von Diskussionen in Arbeitsforen zu extrahieren. Die Literatur diskutiert Probleme bei der Aufgabenauswertung als am weitesten verbreitet, während die Daten darauf hindeuten, dass schlechte Aufgabengestaltung durch Auftraggeber das problematischste Thema für die Arbeitnehmer in den Crowdsourcing-Prozessen zu sein scheint. Frühere Forschung zeigt, dass unerfahrene Auftraggeber keine klare und vollständige Aufgabenbeschreibungen, was direkt zu minderwertigen Arbeitsergebnissen führt. In dieser Arbeit zielen wir darauf ab, diesem Problem zu begegnen, indem wir untersuchen, ob eine automatisierte Schreibhilfe Auftraggebern ermöglichen kann, Unklarheiten in ihren Aufgabenbeschreibungen vor der Veröffentlichung auf der Plattform zu identifizieren und zu verbessern. Um dies zu erreichen, unternehmen wir zwei wichtige Maßnahmen: a) Wir untersuchen zunächst, ob Unklarheiten in Aufgabenbeschreibungen mithilfe von Methoden zur natürlichen Sprachverarbeitung automatisch identifiziert werden können. Wir identifizieren und synthetisieren acht Unklarheiten in Aufgabenbeschreibungen und erstellen sowohl BERT-basierte als auch merkmalsbasierte binäre Klassifikatoren. In einer Crowdsourcing-Studie sammeln wir Anmerkungen zu Unklarheiten in 1332 realen Aufgabenbeschreibungen. Anhand dieses Datensatzes bewerten wir verschiedene Konfigurationen der Klassifikatoren. Unsere Ergebnisse zeigen, dass nahezu alle Unklarheiten in Aufgabenbeschreibungen von den Klassifikatoren recht gut bewertet werden können. b) Aufgrund der Erkenntnisse aus dem vorherigen Schritt haben wir ein Tool entwickelt, das Auftraggebern ermöglicht, iterativ acht häufig auftretende Unklarheiten in ihren Aufgabenbeschreibungen vor der Veröffentlichung auf Plattformen zu identifizieren und zu korrigieren. Es verwendet Modelle zur natürlichen Sprachverarbeitung, die an realen Aufgabenbeschreibungen trainiert sind und eine gegebene Aufgabenbeschreibung auf die acht Unklarheiten bewerten. In einer zweiphasigen Benutzerstudie bewerten wir, ob die automatisierte Unterstützung bei der Erstellung von Aufgabenbeschreibungen aus

Sicht der Auftraggeber nützlich ist und aus der Perspektive der Crowdarbeiter, die in der Praxis mit solchen Beschreibungen konfrontiert sind, effektiv ist. Basierend auf unseren Ergebnissen bewerteten etwa 65% der Auftraggeber die Informationsunterstützung des Tools als sehr hilfreich oder sehr hilfreich. Darüber hinaus berichteten 76% der Crowdarbeiter von einer Verbesserung der Gesamtqualität der Aufgabenbeschreibungen, wenn die Auftraggeber das Tool nutzten. Die Ergebnisse deuten darauf hin, dass wir mithilfe von Techniken zur natürlichen Sprachverarbeitung Auftraggebern automatisch bei der Identifizierung von Unklarheiten in ihren Aufgabenbeschreibungen unterstützen können, was zu Verbesserungen führt, die sie für Arbeitnehmer klarer machen. Dies wiederum führt zu einer verbesserten Aufgabengestaltungsqualität und löst somit das Problem der Qualität der Einreichungen von Arbeitnehmern als eine Hauptherausforderung in Crowdsourcing-Prozessen.

Acknowledgment

I would like to express my heartfelt gratitude to the individuals and organizations who have played instrumental roles in the successful completion of this Ph.D. journey.

First and foremost, I extend my sincere appreciation to Prof. Dr. Gregor Engels, my boss at Computer Science department at Paderborn University, for his constant support and understanding. His encouragement, feedback, and flexibility have provided me with the necessary environment to balance my professional commitments while pursuing academic excellence.

I am also deeply thankful to my first supervisor, Prof. Dr. Henning Wachsmuth, for his unwavering guidance, mentorship, and invaluable insights throughout this research endeavor. His expertise, dedication, and continuous encouragement have been pivotal in shaping the trajectory of this thesis.

I would also like to extend my heartfelt appreciation to Prof. Dr. Ir. Ujwal Gadiraju who collaborated closely with my first supervisor throughout the course of this research. Their combined guidance, expertise, and collaborative efforts have enriched the depth and breadth of this thesis. Their willingness to share insights, engage in constructive discussions, and provide diverse perspectives have significantly contributed to the overall quality of this work. I am truly grateful for their dedication, time, and contributions, which have expanded the horizons of my research and shaped its outcomes in profound ways.

Furthermore, I wish to acknowledge the Digital Future program, which has provided essential financial support for this research. The backing of this program has not only eased the financial burdens associated with academic pursuits but has also enabled me to focus wholeheartedly on my studies and research objectives.

The culmination of this Ph.D. thesis would not have been possible without the steadfast support and contributions of these remarkable individuals and institutions. I am truly honored and humbled by their involvement in this significant achievement.

Chapter 1

Introduction

In this thesis, our objective is to investigate the primary challenges hindering the success of crowdsourcing processes and implement a strategy to tackle the most significant impediment. Initially, we conduct a two-pronged study comprising a "literature review" and an empirical "data analysis" to pinpoint the foremost obstacle exerting the most substantial impact on these processes.

Our findings reveal that the poor task design, particularly the lack of clarity in task instructions provided by requesters has been highlighted as the primary issue. Unclear task descriptions has the most pronounced influence on the quality of submissions and, consequently, the overall quality of crowdsourcing processes. To address this problem, we undertake a computational assessment to determine whether Natural Language Processing (NLP) techniques can yield models capable of automatically detecting clarity flaws of crowdsourcing task instructions.

Given that NLP methods allow us to train models for recognizing clarity issues in task instructions, our aim is to mitigate this challenge by investigating whether an automated writing assistant tool can assist requesters in identifying and enhancing the clarity of their task descriptions (i.e., task's title and the body containing instructions) before deploying them on the platform. We evaluate the tool's effectiveness in assisting requesters to improve the clarity of their task instructions, considering the perspectives of both requesters and workers.

In this chapter, we provide a comprehensive overview of the thesis, encompassing a concise introduction to the field of crowdsourcing, an exploration of the challenges from both theoretical and practical viewpoints , the research questions addressed, and the key contributions made to advance the state-of-the-art in improving crowdsourcing processes.

1.1 Crowdsourcing

The rapid expansion of web technologies and the vast number of internet users have driven the emergence of innovative co-creation models, which integrate external sources of innovation for value creation and problem-solving. One such co-

creation model is crowdsourcing, which was first introduced by Howe et al. (2006) in Wired magazine, inspired by the growing potential for large-scale collaboration facilitated by the internet.

Howe et al. (2006) characterizes the concept of crowdsourcing as a method of addressing problems by outsourcing human-intelligence tasks—typically executed by specific employees—to a group of undefined remote web workers through open calls. This definition highlights crowdsourcing as a transformative work model that has shifted conventional organizational workflows where value creation was primarily performed by in-house employees. Crowdsourcing, instead, serves as a conduit connecting producers and users of company services, enabling volunteers to contribute their innovative ideas and expertise to product development.

Crowdsourcing grants businesses access to an extensive and diverse pool of talents and creativity through web-based platforms. This approach has garnered the interest of companies and organizations across numerous fields, including information management (such as Wikipedia), business and marketing, environmental sciences, medicine, sociology, computer science, and beyond (Hosseini et al., 2014).

In academia, crowdsourcing has evolved into a vital instrument that enables researchers to connect with a broader community, thus expediting the progress of knowledge acquisition. It has left a profound mark on research by harnessing the combined wisdom and assets of a diverse array of individuals, manifesting its influence in various aspects such as swift and economical data collection, fostering interdisciplinary cooperation among specialists from varied disciplines, facilitating rapid dissemination of research findings through peer review, providing essential funding and support for research endeavors, empowering efficient data analysis, and more (Hedges and Dunn, 2017).

Numerous researchers have explored various aspects of crowdsourcing, such as models, applications, workflows, benefits, and challenges. Among these studies, Estellés-Arolas and González-Ladrón-de Guevara (2012) examined crowdsourcing definitions to identify common elements, while Hetmank (2013) investigated typical design aspects of crowdsourcing systems. Hosseini et al. (2014) discussed the four pillars of crowdsourcing and deduced a taxonomy, and Hossain and Kauranen (2015) reviewed the development of crowdsourcing literature and listed its applications. Additionally, Nassar and Karray (2019) summarized methods used in crowdsourcing processing steps, and Muhdi et al. (2011) employed a specific research design to examine the main phases of intermediary-mediated crowdsourcing processes in ongoing projects.

This work forms a part of the "Digital Future" research program[1], an interdisciplinary collaboration between Paderborn University and Bielefeld University, featuring a team of psychologists, sociologists, engineers, economists, and com-

[1]The program's url: https://www.uni-paderborn.de/en/news-item/91403

puter scientists. The primary objective of the Digital Future program is to comprehend and enhance crowdsourcing processes while developing technological methods that assist people—including employers, freelancers, and individuals—in both their professional and personal lives.

To refine the scope of our focus, we concentrate on the overall concept of the general crowdsourcing process. In Chapter 2, we provide a concise overview of the fundamental components of crowdsourcing processes, including tasks, requesters, crowd workers, submissions, and intermediary platforms.

In short, in the general crowdsourcing process (Howe et al., 2006), "tasks" serve as the core components, representing the assignments that require completion through crowdsourcing. These tasks are defined and incentivized by "requesters," who can be either individuals or organizations, initiating crowdsourcing endeavors. On the other end, contributors or participants known as "crowd workers" are the individuals responsible for executing the tasks outlined by requesters. These crowd workers constitute a diverse and geographically dispersed online community. Once completed, the work carried out by these crowd workers is submitted. These "submissions" play a pivotal role in the crowdsourcing ecosystem as they are subsequently assessed and employed by requesters for their intended objectives. "Intermediary platforms" (Howe et al., 2006) play a crucial role in this process, functioning as online facilitators that mediate the interactions between requesters and crowd workers. Some notable examples of such platforms include Amazon Mechanical Turk (MTurk)[2] and Upwork[3].

The subsequent sections provide an overview of the thesis development process, presenting the focus of this work and the central research questions to be addressed.

1.2 Problem Statement and Approach

As explained in Section 1.1, the crowdsourcing model can be regarded as a co-creation approach that harnesses the collective intelligence of a diverse group through an open call, leading to the realization of collaborative services and innovative ideas in a more cost-effective and expedited manner compared to the conventional employment model.

Various areas such as information systems and human computation, psychology, business, and organization management have adapted the crowdsourcing processes, and a wide range of requesters with various backgrounds actively benefit from the merits of the models. Crowdsourcing platforms are necessary for connecting the extensive network of requesters and workers spanning diverse cultures,

[2]MTurk's homepage link: `https://www.mturk.com`

[3]Upwork's homepage link: `https://www.upwork.com`

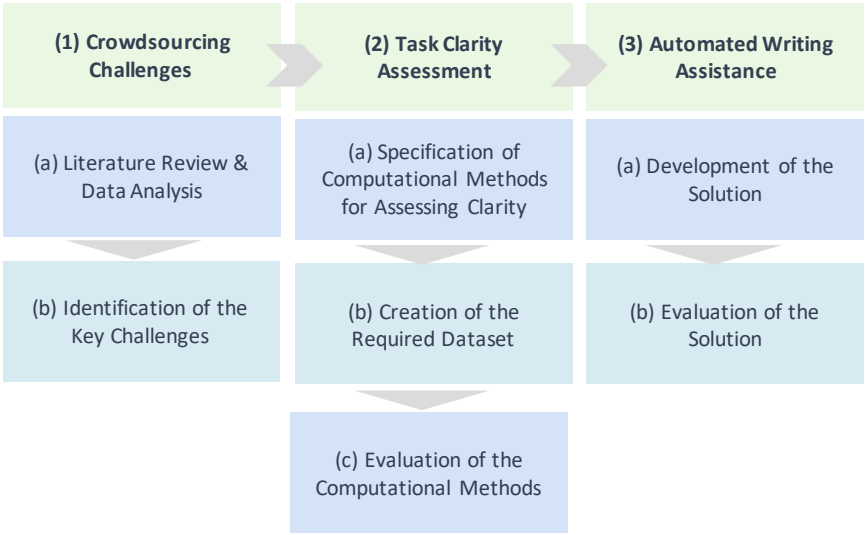


FIGURE 1.1: Overview of the thesis process

possessing different skills, and varying educational backgrounds from around the globe in an anonymous and distant setting.

According to the promises made by crowdsourcing models, on the one hand, requesters profit from crowdsourcing models by obtaining high-quality submissions. On the other hand, workers pursue monetary or non-monetary incentives such as reputations or skills. Their motivation fulfillment relies on acceptance of their submissions, remuneration, or other specified rewards. In practice, however, crowdsourcing models are not entirely successful in fulfilling their promises to both requesters and crowd workers.

In line with the primary goals of the "Digital Future" research program, our initial focus was on exploring the impediments that hinder the effectiveness of crowdsourcing models. This investigation allowed us to gain a thorough understanding of the various challenges present at each stage of the process, potentially leading to the discovery of underlying relationships and the reasons behind these challenges. By obtaining this comprehensive view, we are equipped to make significant contributions to the state-of-the-art, focusing on the development of beneficial solutions for requesters, crowd workers, and ultimately, the overall success of crowdsourcing models. Hence, we adhered to a process to achieve this ultimate goal.

Figure 1.1 provides an overview of the developmental process followed throughout this thesis. The process consists of three primary sequential phases, which are listed and elucidated below:

1. ***Crowdsourcing Challenges***: We conduct a twofold study to obtain a broad overview of the crowdsourcing challenges and their dominance (i.e., *(a) Literature Review & Data Analysis*). Next, we analyze the prior study's find-

ings and identify the major challenges including research questions (i.e., *(b) Identification of the Key Challenges*).

2. ***Task Clarity Assessment***: We aim to adopt an approach of leveraging computational techniques (i.e., *(a) Specification of Computational Methods for Assessing Clarity*) to study whether task clarity as one of the key challenges can be computationally assessed. On this basis, we specify the fundamental components and create the required corpus for the assessment study (i.e., *(b) Creation of the Required Dataset*). We then evaluate whether natural language processing techniques could provide effective models for the problem we focus on addressing in this thesis (i.e., *(c) Evaluation of the Computational Methods*).

3. ***Automated Writing Assistance***: Building upon the findings from the assessment study in the previous step, we proceed to develop the necessary models. Subsequently, we employ these models to create an automated interactive tool to investigate whether these models and such an assistant tool can address the problem (i.e., *(a) Development of the Solution*). To assess the effectiveness of our solution, we evaluate our tool against our main objectives (i.e., *(b) Evaluation of the Solution*).

In Step 1.a, titled "Literature Review & Data Analysis" (Fig. 1.1), we carry out a comprehensive study to gather both widely recognized and less acknowledged challenges. This approach allows us to gain a comprehensive understanding of the existing issues and their significance in crowdsourcing processes, providing valuable insights into areas that necessitate improvements.

Our main emphasis is on investigating the challenges faced by crowd workers, as they constitute the largest proportion of human actors in crowdsourcing and are the primary source of creative services and innovative ideas within the crowdsourcing process. To accomplish this, we delve into the literature on crowdsourcing challenges, primarily adopting a theoretical perspective, and also explore crowd workers' forums, where workers share their daily experiences and the challenges they encounter during the process. Our research primarily revolve around addressing two key research questions:

- **RQ1:** What problems do workers face in the different phases of crowdsourcing processes?

- **RQ2:** Which of these problems are most dominant in the literature and the data, respectively?

In Step 1.b, titled "Identification of the Key Challenges" (Fig. 1.1), we conduct an analysis of the previous study's findings, revealing that the most significant issue

causing additional challenges for crowd workers in crowdsourcing processes is the failure of requesters to design fair tasks in terms of time and effort estimation, as well as providing clear instructions. While prior research, such as the work by Kittur et al. (2013), primarily emphasized low-quality submissions from crowd workers as the central challenge in crowdsourcing, an extensive body of literature highlights that the clarity of task design by requesters plays a pivotal role in shaping the quality of workers' submissions (Khanna et al., 2010, Chandler et al., 2013, Gadiraju et al., 2017, Gaikwad et al., 2017, Wu and Quinn, 2017, Nouri et al., 2020).

In general, a task description should be easy to understand and follow (Alagarai Sampath et al., 2014) with clear terms, and also should describe sufficiently what is expected to be delivered by workers and how this should be done (Kittur et al., 2008, Grady and Lease, 2010, Alonso and Baeza-Yates, 2011, Franklin et al., 2011, Finnerty et al., 2013, Manam et al., 2019).

The failure of requesters to create clear task descriptions can be attributed to two main complexities: First, requesters are responsible for crafting task descriptions that include all essential information, such as the required resources, step-by-step instructions, and the format for submitting solutions. This responsibility becomes particularly burdensome for those without extensive crowdsourcing experience, especially when dealing with micro-tasks that appeal to a diverse pool of workers with varying skills and educational backgrounds (Ipeirotis, 2010). Second complexity pertains to the challenge of expressing task descriptions clearly and comprehensibly arises due to the inherent ambiguity of natural language and the subjective language used by requesters. Consequently, workers may interpret the instructions and requirements differently, leading to potential discrepancies in their submissions (Franklin et al., 2011).

Hence, in Step 2.a, titled "Specification of Computational Methods for Assessing Clarity" (Fig. 1.1), we propose addressing the dual challenge of equipping requesters with essential information presented in a clear manner. This calls for the utilization of computational techniques to aid requesters automatically, enabling them to enhance the clarity of their task descriptions while ensuring they attain the required level of completeness.

In order to explore the concept of an automated solution that assists requesters in identifying clarity issues in their task descriptions, we need to define the criteria for a clear crowdsourcing task description. Subsequently, we proceed to create a dataset, as shown in Step 2.b titled "Creation of the Required Dataset" (Fig. 1.1), which are used to train computational models for this purpose.

In Step 2.c, titled "Evaluation of the Computational Methods" (Fig. 1.1), our objective is to evaluate the extent to which task description clarity flaws could be identified computationally through natural language processing techniques applied

to their plain text. Therefore, we focus on addressing the following two research questions:

- **RQ3.** How effectively can clarity flaws in task descriptions be identified automatically?
- **RQ4.** What textual properties render a task description unclear concerning the defined flaws?

Given the insightful outcome gained from the previous study, which indicate that computational techniques can effectively enable automatic assessment of clarity regarding defined flaws, our focus shifts to Step 3.a, titled "Development of the Solution" (Fig. 1.1). In this step, our objective is to tackle the task clarity problem by constructing models with the ability to assess the level of clarity in task descriptions. To evaluate the effectiveness of our models, we utilized them to build a web-based tool designed to aid requesters in writing clearer task descriptions.

Following the development of the tool, we then proceed to Step 3.b, titled "Evaluation of the Solution" (Fig. 1.1), where we evaluate its effectiveness based on both requesters' and crowd workers' judgments. During this evaluation, we investigate the following two primary research questions:

- **RQ5.** How effectively can an assistant tool help requesters to identify the clarity flaws in task descriptions?
- **RQ6.** How effectively can such a tool help to create task descriptions clearer to workers?

In the upcoming sections, we will provide brief explanations of each step undertaken in this work, which highlights our contributions in tackling the issue of unclear task descriptions in crowdsourcing. By enhancing workers' understanding of task instructions, our efforts are geared towards potentially increasing the satisfaction of both requesters and crowd workers, ultimately contributing to the overall success of crowdsourcing marketplace.

1.3 Crowdsourcing Challenges

In *Literature Review & Data Analysis* step (Fig. 1.1(1.a)), we conduct a comprehensive survey to obtain insights into the underlying reasons for the barriers to the full success of crowdsourcing processes. In particular, we apply two complementary methods:

- **Literature review**: We collect the challenges discussed in the literature, mostly found from interviews with workers. This source of information provides us with an overall view of the challenges uncovered by researchers.

- **Empirical data analysis**: We mine challenges from the complaints that workers shared with their community about the complications, confusions, and unfairness they face during work. We hypothesize that discussions in crowd workers' forums contain the problems and their significance in practice that researchers may not have identified yet.

By adopting this approach, we gain a deeper understanding of the existing challenges associated with crowdsourcing processes, serving as the foundation for future research. Based on this hypothesis, we direct our attention to two primary research questions: RQ1, which involves exploring challenges through literature and crowd worker's forums, and RQ2, which aims to analyze the dominance of these obstacles. In the subsequent sections, we will provide brief explanations of each method used and the corresponding data collected separately. Finally, we will conclude by summarizing the findings derived from our investigation.

Literature Review

To collect the challenges discussed in articles, we conduct a comprehensive literature survey, reviewing studies that specifically focused on the challenges faced by crowd workers throughout the processes. These challenges were sourced from surveys and, in part, from interviews conducted with crowd workers.

Literature We select relevant articles encompassed crowdsourcing problems approached from questionnaires, surveys on crowdsourcing platforms, or face-to-face interviews. In line with the data used for data analysis, the majority of studies focused on Amazon Mechanical Turk (MTurk). MTurk is widely recognized as the largest and most popular micro-task crowdsourcing platform (O'Neill and Martin, 2013).

Method Through manual analysis, we identify and categorize the problems based on their relevance to either requesters' or workers' performance and classified them according to the different phases of the crowdsourcing process.

Empirical Data Analysis

As a supplementary approach to the literature review, we employ topic modeling to identify the most commonly discussed problems among workers on an online discussion forum, drawing from their firsthand experiences in crowdsourcing processes. Through this method, we analyze the narratives that include genuine complaints about the difficulties they encounter while working for requesters.

This section serves as a summary of both our analysis and the corpus that we have established for this research and potential future studies within the crowdsourcing marketplace.

Data Over time, various worker community forums have been established to facilitate communication and mutual support among workers during crowdsourcing processes. Among these forums, we specifically analyze the data from the *Turkopticon* forum, as it is the most widely used platform for sharing daily crowdsourcing stories (Irani and Silberman, 2013). For the corpus of our data analysis, we performed a crawl of all stories (i.e., reviews about requesters) on Turkopticon, specifically focusing on the negative experiences, as they encompass the various challenges workers encounter during their work.

Method In the data analysis process, our objective is to computationally mine crowdsourcing problems from the reviews. To achieve this, we employ the Latent Dirichlet Allocation (LDA) method (Blei et al., 2003), which identifies hidden topics among a large set of documents in a corpus.

In our approach, we treat each individual review as a separate document, and through LDA, we extract the primary problems (i.e., topics) mentioned in the reviews.

Results

In this stage, we conduct a comparison between the findings derived from the data analysis and those obtained through the literature review process. This allows us to gather and juxtapose the challenges discussed among workers in practice with the perspectives presented by researchers from a more theoretical viewpoint.

In light of RQ1, we identify a total of 14 distinct challenges discussed both in the literature and among workers. The results indicate that these challenges manifest across all stages of the crowdsourcing process. However, the main issues experienced by workers are primarily related to task design, such as encountering vague task descriptions and facing underestimations. Additionally, errors in the task environment lead workers to expend time on unsuccessful submissions. Lack of feedback and inadequate responses to workers' inquiries pose significant problems during the task operation phase. The literature also extensively covers communication challenges between requesters and workers, often compounded by poor platform support. Simultaneously, unfair rejections without proper explanation emerge as the dominant problem during the task evaluation phase. In light of RQ2, we can infer that the challenges originating in the task design stage have the most significant impact, subsequently leading to issues in the subsequent stages of the crowdsourcing process. For instance, poor task design can result in low-quality

submissions from workers and unfair rejections by requesters, subsequently leading to poor communication between workers and requesters due to the lack of justifiable reasons for rejections. During the task design step, requesters are expected to accurately estimate the required time, effort, and fair payment for their tasks. Additionally, they should provide clear and concise instructions to facilitate the submission of desired solutions, however, they often struggle to meet this requirement effectively in practice.

The results obtained in this section have unveiled a wide array of potential directions for further research aimed at enhancing crowdsourcing processes. In the following section, we will delve into the focus of our study, informed by the findings from this section.

1.4 Task Clarity Assessment

Based on the insights gained from the previous study, it becomes evident that challenges stemming from the task design stage have the most substantial impact on the subsequent task operation and evaluation stages. Additionally, the study revealed that unclear task descriptions have a considerable influence on obtaining low-quality submissions, which is identified as the most significant challenge influencing the requesters' satisfaction in the crowdsourcing marketplace.

Clarity of task description mainly relates to the completeness and understandable wording of the instructions. These instructions should furnish all essential information for crowd workers to assess their interest in performing the task and submitting a solution. Consequently, a lack of required knowledge and the presence of ambiguous language may impact how crowd workers perceive the task, potentially leading to reduced participation or low-quality submissions.

For example, a real-world task description written as the follows (For detailed explanation, refer to Section 2.1):

> **Title:** Do a google search
> **Body:** Do a google search to make sure site is indexed

It requires a brief explanation of the term "indexed," which is a crucial aspect of the description, as well as the specific format in which the results of the work should be specified and submitted. The absence of such clarifications in the task description may contribute to ambiguity and hinder workers' ability to complete the task effectively.

We hypothesize that a web-based tool employing natural language processing techniques is a possible solution that may help requesters identify and improve clarity flaws in their task description. To validate this hypothesis, our initial focus is on investigating the feasibility of computationally assessing clarity flaws in task

descriptions. To achieve this, we design two preliminary steps, as shown in Figure 1.1(2):

Computational Methods for Assessing Task Clarity

Upon achieving the objectives of the first part of the process (Fig. 1.1(1)), we aim to address two research questions in the next step of the process (Fig. 1.1(2.c)): the extent to which computational assessments of task description clarity are effective (RQ3), and what textual properties of the descriptions are significant for clarity flaw evaluations by models (RQ4). To address the two research questions, we investigate two approaches namely, a state-of-the-art neural model and b) a traditional feature-based model. We employ both methods to compare the efficacy of neural models with that of feature-based models in classifying the clarity of task descriptions. In order to develop feature-based models, we need to apply our domain-specific expertise on crowdsourcing task descriptions to create feature types. To this end, we design six distinct feature types: *content*, *length*, *style*, *subjectivity*, *readability*, and *flaw-specific* features, which capture various aspects of task descriptions.

Creation of the Required Dataset

In the preliminary steps of investigating computational methods, we carry out fundamental preparations, including the identification of explicit dimensions related to crowdsourcing task description clarity. Following the literature discussing dimensions of unclear task descriptions, we form a set of clarity dimensions of crowdsourcing task descriptions relating to comprehensibility and completeness. To train models, we also create the necessary corpus of annotated task descriptions with respect to the defined task clarity dimensions, encompassing both clear and unclear task descriptions. For the annotation, we rely on the judgment of crowd workers regarding the clarity of the task descriptions in the dataset. These trained models enable us to predict and identify clarity flaws in task descriptions.

Results

Regarding RQ3, both approaches demonstrate learning success in nearly all cases, with the exception of identifying a difficult wording and phrasing. We observe that the baseline ranges from 0.31 to 0.71, while the BERT models exceed the baseline with results ranging from 0.55 to 0.71. and the Support Vector Classifiers (SVCs) outperform BERT, achieving results between 0.61 to 0.74 for the majority of clarity flaws. Regarding RQ4, we observe that the content, style, and readability seem to be significant textual properties for clarity. Combining the task flaw-specific prop-

erties with others is also advantageous for clarity assessment.

In this study, the superior performance of SVCs demonstrates that task clarity can be computationally assessed using the features we defined based on domain-specific knowledge. Consequently, we employ these features to build models capable of predicting clarity flaws in task descriptions. By deploying these models, we create and evaluate a web-based tool that serves as an assistant system for automatically helping crowdsourcing task requesters identify potential clarity flaws in their task descriptions. This tool can guide requesters in determining the information they need to include in their task descriptions to enhance clarity. The development of the tool are briefly described in the following section.

1.5 Automated Writing Assistance

Drawing on insights from the assessment study of the computational approach, we envision an interactive tool that automatically analyzes the plain text of a task description and predicts its potential clarity flaws as a valuable solution. However, to the best of our knowledge, such a tool has yet to be created, given the absence of practical computational methods to assess task description clarity. To address RQ5 and RQ6, we develop the solution in two primary steps:

- **Solution Development**: We construct the necessary models that enable the development of a web-based tool which requesters can utilize to iteratively evaluate the clarity of their task descriptions and identify the defined clarity flaws. This tool allows requesters to enhance their description clarity before posting it on the crowdsourcing platform.

- **Solution Evaluation**: We carry out a twofold evaluation study involving requesters and workers to examine the tool's helpfulness for requesters, as well as its effectiveness in resolving unclear task description issues for workers.

In the following, we provide a brief introduction to our tool and we discuss the user studies conducted to evaluate the tool's effectiveness based on the judgment of requesters and crowd workers.

ClarifyIt: A Writing Assistance Tool for Task Descriptions

To address the issue of unclear task descriptions, we develop a tool called *ClarifyIt* (where 'It' refers to both the task description and the iterative process) that functions as an assistant system for requesters to either create clear crowdsourcing task descriptions from scratch or identify potentially unclear sections in their

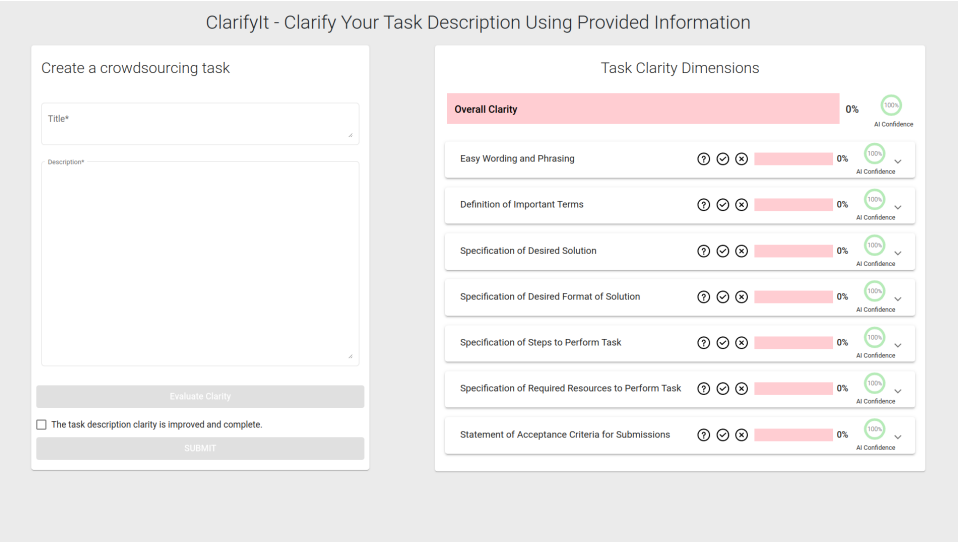


FIGURE 1.2: User interface of our automated writing assistance tool (*ClarifyIt*)

existing task descriptions. Using the tool, requesters can find and edit known ambiguities or incomplete information in their task descriptions through an iterative process before deploying them on the platform.

We utilize natural language processing techniques to develop ClarifyIt, crafting computational models that automatically analyze task descriptions to identify pre-defined clarity flaws. We envision that crowd workers, who work on tasks for requesters using ClarifyIt, will receive clearer descriptions in terms of completeness and understandable phrasing.

Our tool serves two essential functions in assisting requesters: (a) It supplies inexperienced requesters with clarity dimensions that should be considered to minimize task ambiguity. (b) It offers scores reflecting the clarity level of the description based on the pre-defined clarity dimensions.

In order to develop ClarifyIt, we initially need to adjust the dataset, train appropriate models, and subsequently implement the tool. Figure 1.2 illustrates the tool's user interface, allowing requesters to input the task title and description on the left side while presenting clarity scores for the description on the right side. A comprehensive exploration of the tool's user interface will be provided in Chapter 5.

Evaluation of ClarifyIt

We evaluate the tool's helpfulness and effectiveness in enhancing task description clarity through a process that involves both requesters and workers. Specifically, we aim to determine how effectively ClarifyIt assists requesters in identifying clar-

ity flaws (RQ5) and how well it contributes to the creation of clearer task descriptions that benefit crowd workers (RQ6).

Therefore, the evaluation process of ClarifyIt comprises two major sequential steps. (a) *Evaluation with requesters* - We recruit requesters to use our tool to create task descriptions and iteratively improve their clarity. The requester generates the first version of the description, iteratively assesses and enhances its clarity using ClarifyIt, and finally submits all versions of the task descriptions. (b) *Evaluation with crowd workers* - We ask workers to compare the texts of two versions of the task descriptions and provide their opinion on the clarity improvements of the task descriptions created in the previous step.

Results

The evaluation results reveal that 65% of all requesters found our tool helpful in terms of its functionalities and the information provided on task description clarity dimensions (RQ5). Only 12% held the opposite opinion, primarily due to the quality of examples shown in the tool and the accuracy of prediction models. Furthermore, 60%–78% of the crowd workers agreed that all clarity dimensions improved in the task descriptions edited version using our tool (RQ6). The results also suggest that our tool is most effective in clearly defining the desired solution for tasks within the instructions. However, providing automated support for improvements in precise and straightforward wording and phrasing of task descriptions is more sophisticated compared to other clarity dimensions.

While our tool's effectiveness can be enhanced in some respects, we conclude that, in addition to the prior theoretical contributions classifying ambiguous crowdsourcing task descriptions, our tool can assist requesters in improving their task description clarity in practice without requiring crowd workers' involvement. In future research, it would be worthwhile to explore a similar approach geared towards supporting content writers in enhancing the clarity of their text, based on the crucial dimensions of text clarity in their domain.

1.6 Overview of Publications

In this section, we provide a summary of the key milestones and related findings in our thesis, which are documented in various publications. We have published three complete papers and one Work-in-Progress (WIP) paper, the contents of which are reused and reported in this thesis. Below, we provide a brief overview of these publications.

Mining Crowdsourcing Problems from Discussion Forums of Workers ((Nouri et al., 2020) - COLING 2020) To better understand the crowdsourcing domain and identify areas needing improvement, we initially carried out a twofold study to gather the challenges faced by workers in various crowdsourcing processes, given the significance of their role in the success of crowdsourcing marketplaces. Our approach and findings are elaborated in (Nouri et al., 2020), where we explored the mentioned challenges and their prominence from both theoretical and practical perspectives in crowdsourcing workflows. This study offered a comprehensive understanding of potential areas for enhancement and viable directions for future research.

What is Unclear? Computational Assessment of Task Clarity in Crowdsourcing ((Nouri et al., 2021a) - HT 2021) Upon analyzing insights from previous research on worker challenges, we determined that the clarity of task instructions provided by requesters during the task design phase considerably influences the quality of workers' submissions. We thus concluded that unclear task descriptions contribute significantly to the existing challenges and subsequently have a considerable impact on workers' satisfaction. As a result, we concentrated on enhancing the clarity of crowdsourcing task descriptions to improve crowdsourcing processes and ultimately contribute to the success of crowdsourcing marketplaces. In general, our aim was to employ computational methods to assist both experienced and novice requesters in writing clear task descriptions prior to posting them on a platform. To realize this vision, we first needed to investigate whether natural language processing techniques offer us efficient technological tools (i.e., models) capable of evaluating task description clarity based solely on their plain text. In (Nouri et al., 2021a), we presented our threefold contribution which includes: the creation of a necessary corpus for the computational evaluation of task description clarity, a feature-based and a neural approach for assessing task clarity, and empirical insights into crucial aspects of computational assessment of task clarity. In the paper, we detailed our work's focus, approach, and insightful findings extensively.

iClarify – A Tool to Help Requesters Iteratively Improve Task Descriptions in Crowdsourcing ((Nouri et al., 2021b) - HCOMP 2021) The computational assessment of unclear task instructions demonstrated that natural language processing techniques furnish us with efficient models for predicting clarity issues in task instructions. Consequently, we developed a web-based tool (called *iClarify*[4] - Iteratively Clarify) to help requesters identify clarity flaws in their task descriptions and iteratively improve them until a satisfactory level of clarity is achieved. In (Nouri et al., 2021b), we introduced the concept of an assistant tool for requesters to gener-

[4]We later changed the tool's name to ClarifyIt to prevent violating the rules of the blind review process.

ate and enhance their task description clarity, showcasing the initial version of the tool prior to evaluation studies in the form of Work-in-Progress (WIP) research.

Supporting Requesters in Writing Clear Crowdsourcing Task Descriptions Through Computational Flaw Assessment ((Nouri et al., 2023) - IUI 2023) Relying on the results in (Nouri et al., 2021a), we built support vector regression models with various feature types for task clarity flaws that predict the degree of clarity flaws in crowdsourcing task descriptions. Our tool enables requesters to iteratively assess their instruction clarity and improve it until the scores shown by the tool reach sufficient clarity. We conducted two user studies with requesters and crowd workers to evaluate, on the one hand, how well the tool assists requesters in improving their task instructions clarity according to their judgment, and on the other hand, how much the clarity of the instructions improves through our tool in practice according to the workers' judgment. In (Nouri et al., 2023), we thoroughly presented the fundamental aspects of our work, the methodology, and the design of the evaluation studies along with their findings.

1.7 Overview of Thesis Structure

Figure 1.3 shows an overview of the thesis structure. The thesis is divided into six main chapters, namely *Introduction*, *Background*, *Crowdsourcing Challenges*, *Task Clarity Assessment*, *Automated Writing Assistance*, and *Conclusion*.

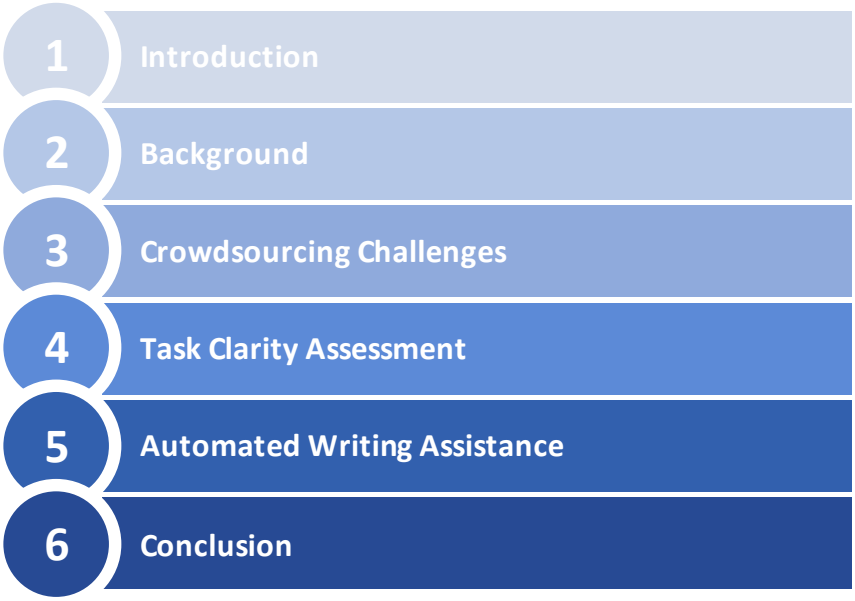


FIGURE 1.3: Overview of thesis structure

The *Introduction* chapter outlines a brief overview of the entire work, introduces the general process of this thesis development, details the research questions, and describes the method employed to address them.

The *Background* chapter provides a thorough review of the relevant literature pertaining to our research focus. Additionally, it offers a concise overview of the methods and algorithms employed to address the research questions outlined in this thesis.

The *Crowdsourcing Challenges* chapter elaborates the method we employed to a comprehensive view on the crowdsourcing problems discussed in theory by researchers and in practice among workers in discussion forums.

The *Task Clarity Assessment* chapter covers the notion of applying computational techniques to develop automated solutions. Specifically, we define crowdsourcing task description clarity and design a method to create the required corpus, enabling us to deploy machine learning methods, especially natural language processing techniques, to tackle the unclear task description problem in crowdsourcing marketplaces. This part also explains the approach we adapt to study whether a computational assessment of task clarity flaws in crowdsourcing is generally feasible. These two steps are preliminary to investigating whether an interactive tool can help requesters identify and improve their task description clarity automatically.

The *Automated Writing Assistance* chapter introduces the computational models and the tool that we developed to target helping requesters to improve the clarity flaws of their task descriptions. Our tool provides an environment where requesters can iteratively evaluate their task description for the known clarity flaws and edit them before posting on a crowdsourcing platform. Moreover, it discusses the twofold user study that we conduct to evaluate to which extent the solution is helpful for requesters and has an impact on clarity improvements of task descriptions in crowdsourcing processes.

The *Conclusion* chapter summarizes the problem we intended to tackle and how effectively we resolved this problem. Furthermore, it offers a glimpse into potential future endeavors, which encompass implementing our solution in different domains and enhancing its effectiveness in bolstering the clarity of crowdsourcing task descriptions.

Chapter 2

Background

This chapter delineates crowdsourcing general process and its essential elements and presents a succinct explanation of the methods and algorithms utilized to address the research questions outlined in this thesis. At last, it offers a comprehensive review of the pertinent literature related to our research focus. In Section 2.1, we delve into the crowdsourcing model and the main stakeholders, elements, and their interrelationships in this process. Moving on to Section 2.2, we offer an overview of Natural Language Processing (NLP) and its diverse applications within the realm of machine learning, along with an exploration of the algorithms that find application in this thesis. Lastly, Section 2.3 provides a detailed exploration of the pertinent prior research aimed clarity assessment and enhancing crowdsourcing processes.

2.1 Crowdsourcing Processes

Crowdsourcing is a novel business model that seeks to mobilize a motivated group of individuals (i.e., crowd workers), capable of providing solutions of superior quality and quantity compared to those achievable through conventional business approaches (Brabham, 2008a). The inclusion of individuals from different time zones opens up the opportunity for business owners (i.e., requesters) to execute projects through an open call on online platforms at any given time, and a substantial workforce ensures the rapid accomplishment of tasks (Berg et al., 2018).

In general, crowdsourcing processes involve a series of phases during which stakeholders (i.e., requesters and crowd workers) or the intermediary platform take actions to accomplish specific goals. The process design can range from simple to sophisticated, depending on the nature of the task (i.e., individual or collaborative contribution) and the domain in which crowdsourcing is being employed (e.g., software development, micro-tasks) (Pedersen et al., 2013). Figure 2.1 highlights three primary phases of the general crowdsourcing process:

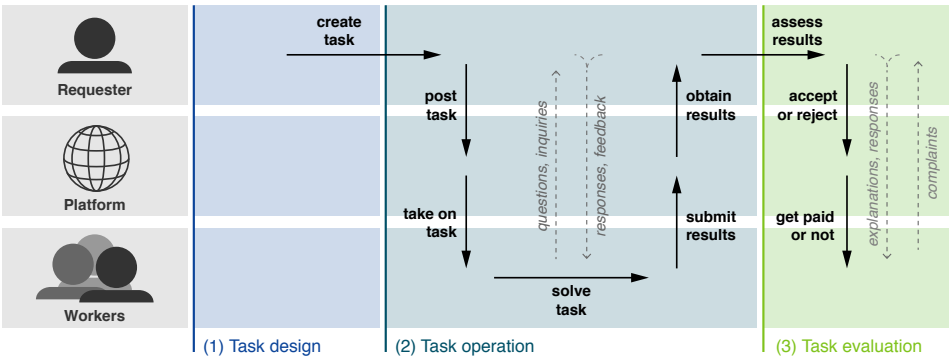


FIGURE 2.1: The three stages of a crowdsourcing process are: *(1) Task design: The requester formulates the task. (2) Task operation:* Workers accept and complete the task, subsequently submitting their results. *(3) Task evaluation:* The requester reviews and either accepts or rejects the results. Communication can occur during both the task execution and evaluation stages (Nouri et al., 2020).

1. *Task design*: The crowdsourcing process begins when requesters define expectations, develop strategies, and design tasks to achieve their desired objectives. During the task design phase, a requester should clearly describe the task and specify the expected solution from workers before posting it on a suitable crowdsourcing platform.

2. *Task operation*: Crowd workers then browse the available tasks on a platform, review the task descriptions, and assess their interest in undertaking the task. If they choose to accept the task, they start working on it and subsequently submit their solution within the specified deadline.

3. *Task evaluation*: Requesters assess the submitted solutions from workers based on their acceptance criteria, determining which submissions to accept and compensate or which to reject due to their insufficient quality in meeting expectations. Workers subsequently receive the results of the requester's evaluation and may respond to any rejections.

During the task operation and evaluation steps, communication between requesters and workers may be necessary. For instance, workers might have questions about payment details, reasons for rejections, or require clarification on the task. Additionally, requesters may offer feedback to help workers enhance the quality of their submissions, thereby increasing the likelihood of being rewarded for their efforts.

In the following, we outline the key elements of the general crowdsourcing process, including *tasks*, *requesters*, *crowd workers*, *submissions*, and *intermediary platforms*, as well as their interrelationships.

Crowdsourcing Tasks

The initial element of the crowdsourcing process is the task, which refers to the job posted by requesters on crowdsourcing platforms, describing the expected deliverables. As outlined by Hosseini et al. (2014), crowdsourcing encompasses a broad array of task types, such as (a) data collection tasks, where the crowd provides information, annotates data, or participates in surveys, (b) innovation tasks, calling for creative designs or ideas, (c) fundraising tasks, necessitating financial services, (d) problem-solving tasks, demanding solutions from the crowd, and (e) co-creation tasks, involving the crowd in the product creation process.

Crowdsourcing tasks can be characterized as (a) Atomic or small units (micro-tasks) as a part of a sophisticated task. (b) They may be simple or complex to solve by human intelligence, yet impossible or extravagant to automate. (c) They may be tasks that crowd workers can accomplish faster with a lower price than hiring experts. (d) They may require individual contributions where workers create solutions alone or a collaborative contribution where a group of workers need to work together to complete a task (Estellés-Arolas and González-Ladrón-de Guevara, 2012, Hosseini et al., 2014).

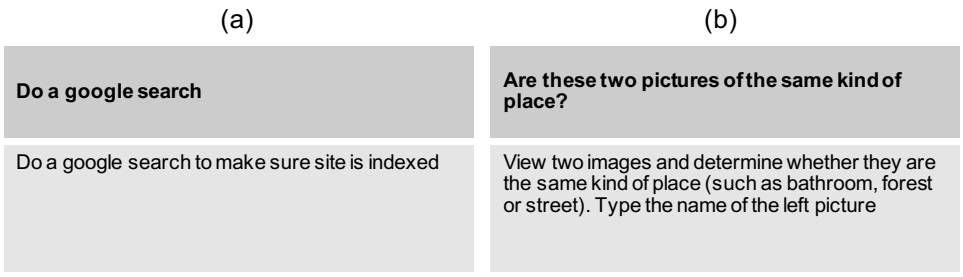


FIGURE 2.2: Two examples of real-world micro-task descriptions on Amazon Mechanical Turk (MTurk) Platform.

Figure 2.2 presents two real-world examples of micro-tasks created by requesters on a crowdsourcing platform for the purpose of data collection. In Figure 2.2(a), the requester asks crowd workers to verify if their website is indexed by Google, while the task in Figure 2.2(b) seeks crowd-sourced evaluation data. In this thesis, the task's title and body, which contain a detailed description written in natural language, are referred to as the *task description*.

Crowdsourcing Requesters

A requester is a stakeholder who engages voluntary online users for tasks such as production, data collection, concept development, and more (Pedersen et al., 2013). Requesters typically include companies, institutions, non-profit organiza-

tions, governments, academic researchers, and private individuals who seek innovative ideas, external knowledge, or additional profit and value. For example, in the task description shown in Figure 2.2(a), the requester might be the owner of the website's company aiming to test how easily users can find their website using the Google search engine. Similarly, the task requester in Figure 2.2(b) might be interested in evaluating the performance of an image processing algorithm that classifies a set of images based on the locations depicted in those pictures.

Crowd Workers

In a comprehensive review of crowdsourcing studies (Estellés-Arolas and González-Ladrón-de Guevara, 2012), crowd workers are defined as online users who voluntarily contribute their skills and knowledge to organizations seeking innovative solutions and services. While crowdsourcing may necessitate well-trained workers (Howe, 2008), crowd workers typically consist of a diverse group ranging from inexperienced amateurs to professional experts or engineers in a given field.

Crowd workers are typically drawn from a vast, unknown network of individuals, except in organized communities where members are more likely to know one another. The size of a potential worker pool for a task depends on factors such as the nature of the task, its confidentiality level, and the qualifications needed for its completion. Some tasks require the insights of a diverse crowd, where personal opinions or knowledge are requested, while others necessitate specific skills, such as text translation or software coding. Consequently, crowd workers represent an undefined group of web users, with the number, diversity, and skill levels of its members contingent on the task requirements. In the task description example shown in Figure 2.2(a), crowd workers from various locations can check if they can access the website through the Google search engine, resulting in a broader range of outcomes that cover a variety of potential cases.

Crowdsourcing Submissions

In the crowdsourcing process, crowd workers must submit their solutions to the task requester after viewing, accepting, and completing the task. As noted in the literature, the outcomes submitted by workers can vary widely depending on the nature of the crowdsourcing tasks, encompassing social feedback, problem resolutions, ideas, talent, external knowledge, skills, experience, added value, increased profit, and product or service innovations (Howe et al., 2006, Howe, 2008, Estellés-Arolas and González-Ladrón-de Guevara, 2012). The quality of these submissions directly impacts the requester's decision to accept the work and, subsequently, compensate the worker. In the task description example depicted in Figure 2.2(a), crowd workers are expected to provide a positive or negative answer regarding whether the Google search engine displays the website in its search results.

Crowdsourcing Platforms

Although there may be scenarios where requesters connect with workers offline or in-person (Howe et al., 2006), requesters typically initiate the crowdsourcing process using online web-based software applications known as crowdsourcing platforms. These platforms connect requesters and crowd workers, mediate the process, and manage the necessary interactions while providing essential utilities. The functionalities of these platforms serve four key purposes (Hosseini et al., 2014):

- Crowd worker side facilities, such as enrollment, authentication, submission, and feedback loop mechanisms
- Requester side facilities, including enrollment, authentication, task posting, result evaluation, and feedback loops
- Task side facilities, like result aggregation, recording worker performance, and qualification filters
- Platform side facilities, encompassing online interfaces, mutual interactions and payment systems.

Examples of such platforms are InnoCentives, Threadless, Amazon Mechanical Turk (MTurk), TopCoder, Crowd4U, Prolific, Upwork, iStockPhoto, CrowdFlower and many more.

Aligned with the core objectives of the "Digital Future" research program, our initial emphasis was directed toward investigating the barriers that impede the effectiveness of crowdsourcing models. By gaining a comprehensive perspective on the existing challenges in the process (Chapter 3), we are well-prepared to make substantial advancements in the field, with a specific focus on leveraging natural language processing techniques to create valuable solutions for requesters and crowd workers, ultimately enhancing the overall effectiveness of crowdsourcing models.

In the following section, we present an introductory overview of the NLP techniques and provide an explanation for its incorporation that underpin our primary research in this thesis. As part of Section 2.3, we provide a summary of the previous research related to the utilization of NLP techniques for assessing clarity.

2.2 Natural Language Processing (NLP)

Computational methods encompass a wide set of methods and techniques employed in the data analysis (Hastie et al., 2009), simulation, modeling, and processing of diverse data types using computers (Fishman, 2013). Machine learning, as a computational method, is a sub-field of artificial intelligence that focuses on

developing models that allow computers to learn and make predictions or decisions without being explicitly programmed (Murphy, 2012). It involves training a computer system to understand patterns and relationships in data, and then using that knowledge to make predictions or take actions (James et al., 2013).

There are various types of machine learning algorithms, including supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves training a model on labeled data, where the desired output is known (James et al., 2013). Unsupervised learning, on the other hand, deals with finding patterns and structures in unlabeled data (Bishop and Nasrabadi, 2006). Reinforcement learning involves training an agent to interact with an environment and learn from the feedback it receives (Sutton and Barto, 2018). Machine learning has a wide range of applications, including image and speech recognition, natural language processing, recommendation systems, fraud detection, and autonomous vehicles. It has revolutionized many industries and continues to drive innovation and advancements in various fields.

Natural Language Processing (NLP) as a sub-field of machine learning is defined as "the study of computational methods for working with human language" (Jurafsky and Martin, 2009). NLP involves developing algorithms and models that enable computers to understand, interpret, and generate human language in a useful and meaningful way. The most known applications of NLP introduced as text classification, named entity recognition, machine translation, question answering, sentiment analysis, and text summarization.

NLP methods are used in this thesis for multiple purposes. In Chapter 3, we apply NLP techniques, specifically topic modeling, to analyze real users' reviews written in natural language to uncover the challenges that they consistently encounter and discuss with others. In Chapter 4, NLP techniques are employed to process textual instructions of crowdsourcing tasks, investigating the feasibility of creating models that classify task descriptions based on their clarity flaws. Likewise, in Chapter 5, our goal is to develop computational models to grade the clarity of crowdsourcing task descriptions in accordance with specific clarity dimensions.

In the following chapters, our objective is to explore particular research questions that involve computational processing of textual data in English natural language, such as reviews written by crowd workers and actual task instructions composed by task requesters. Therefore, in Section 2.2.1, we give a summary on topic modeling technique utilized to analyze reviews and mine challenges faced by workers in their daily work on crowdsourcing platforms (details in Chapter 3). In Section 2.2.2, we briefly describe the Support Vector Machine (SVM) algorithm and its two principal variations, namely Support Vector Classification (SVC) and Support Vector Regression (SVR), while in Section 2.2.5, we provide an overview of BERT models. We utilized SVCs and BERT models to tackle the research ques-

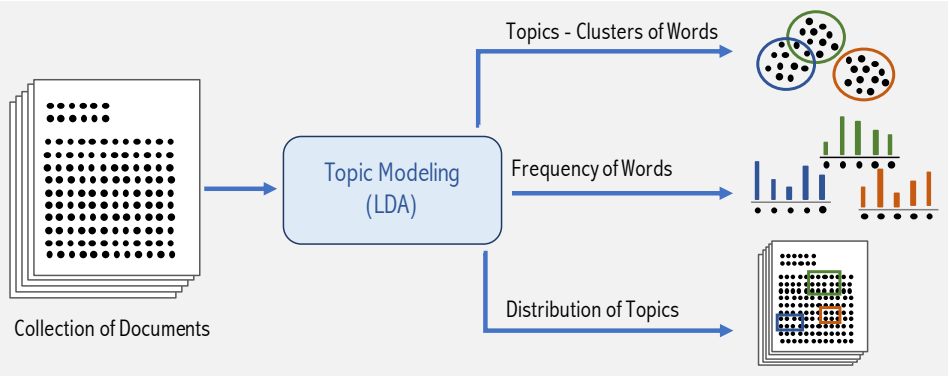


FIGURE 2.3: General overview of topic modeling involves identifying topics and their distributions over each document (Joshi, 2018).

tions in Chapter 4. Meanwhile, we built and employed SVR models in our solution development (elaborated in Chapter 5) to address this thesis's primary goals.

2.2.1 Topic Modeling

Topic modeling is an approach employed in natural language processing and machine learning that automatically uncovers the inherent topics within a document collection. By identifying the primary topics across the documents, this method aids in organizing and comprehending substantial volumes of textual data.

Figure 2.3 provides a general overview of topic modeling where a set of documents is used as input. Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a widely used topic modeling technique, uncovers these hidden topics by analyzing the word patterns across the entire input document collection. Finally, LDA generates the output, consisting of a collection of topics, each characterized by a cluster of words and their respective probabilities. Moreover, each document is linked to one or more topics. This facilitates the comprehension of the primary themes or subjects prevalent within the document collection (Joshi, 2018).

The LDA model has several important parameters that help define its behavior and control the generated results. Some main parameters of LDA are:

- Number of topics (k): This parameter defines how many latent topics the LDA model should identify from the given corpus. Selecting an appropriate value for the number of topics relies heavily on domain knowledge and the particular application.

- Dirichlet priors (α and η): The LDA model uses two Dirichlet priors: one for the document-topic distribution (α) and the other for the topic-word distribution (η). The hyper-parameters α and η determine the shape of these distributions, controlling the diversity of topics in documents and the diver-

sity of words in topics, respectively. Lower values of α or η result in sparser distributions, whereas larger values make them more uniform.

- Max iterations: The LDA model is typically trained using iterative algorithms, and the "max iterations" parameter sets an upper limit for the number of iterations the LDA algorithm undergoes. This parameter influences the convergence of the algorithm and often requires experimentation to establish an optimal value.

A variety of evaluation metrics, including *harmonic mean* (Griffiths and Steyvers, 2004), *pairwise cosine distance* (Cao et al., 2009), and *KL divergence* (Arun et al., 2010) has been proposed to find the best k for a particular corpus. Due to the inaccuracy of these measures, the *chib-style estimator*, which maximizes the probability of held-out documents, was introduced by Wallach et al. (2009b). Nevertheless, the interpretability of topics is not evaluated by this estimator. As a solution, Chang et al. (2009) proposed a metric that measures the topic coherence of models based on human judgments. Newman et al. (2010) also designed a method to assess human judgments, and Mimno et al. (2011), Stevens et al. (2012), and Röder et al. (2015) explored the precision of different coherence metrics.

2.2.2 Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It works by finding a hyperplane that best separates different classes of data points in a high-dimensional space. The key idea behind SVM is to find the hyperplane that maximizes the margin, which is the distance between the hyperplane and the nearest data points of each class. This approach aims to achieve high generalization and robustness to new, unseen data (Cortes and Vapnik, 1995, Schölkopf et al., 2001, Mammone et al., 2009).

SVM's versatility and effectiveness have led to its widespread use in numerous applications such as image classification, text categorization, bio-informatics, and finance due to its strong theoretical foundation and robust performance. It can handle both binary and multi-class classification problems and has been extended for regression tasks as well.

SVMs come in two main variations supporting two main types of machine learning tasks namely classification and regression. In the following, we overview each variation.

2.2.3 Support Vector Classification (SVC)

Support Vector Classification (SVC) is a widely used supervised machine learning algorithm for classification tasks based on the principles of SVM (Vapnik, 1999). The primary goal of SVC is to find an optimal hyperplane that separates the data

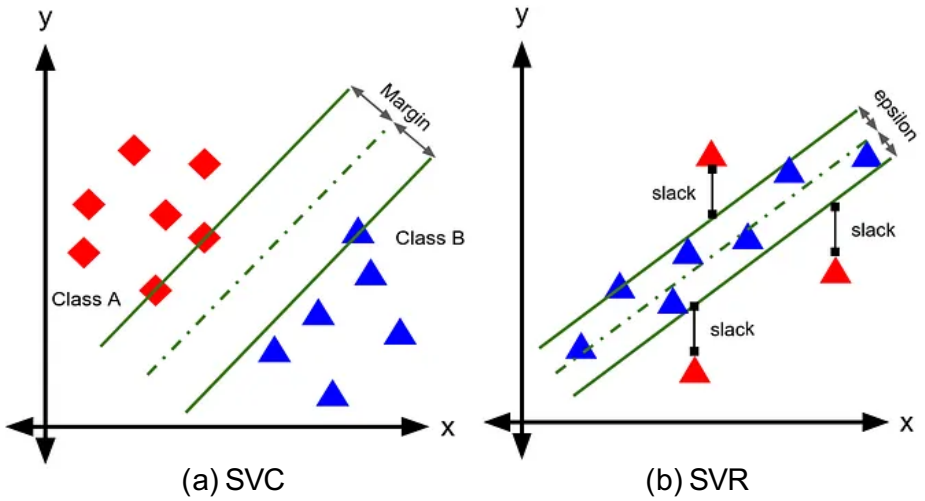


FIGURE 2.4: Demonstration of two variants of a linear Support Vector Machine: (a) Support Vector Classification (SVC of binary version), (b) Support Vector Regression (SVR). The distinction between classification and regression lies in the fact that regression yields a numerical output as opposed to a class. Support vectors represent the data points that are nearest to the hyperplane, and they play a crucial role in improving the definition of the separation line by determining margins. Margins refer to the spaces between the two lines closest to the class points[1].

points belonging to different classes while maximizing the margin between the classes.

As shown in Figure 2.4(a), SVC works by finding a hyperplane that best separates the data points into two classes (in the case of binary classification). The algorithm searches for the hyperplane that maximizes the margin between the two classes, resulting in a high generalization ability. The margin is the distance between the closest data points from each class, known as support vectors, and the separating hyperplane (Vapnik, 1999, Cortes and Vapnik, 1995).

SVC can handle linear and non-linear relationships between features and target classes using kernel functions. Kernel functions map the input data into higher-dimensional spaces, allowing SVC to work in that transformed space and find a hyperplane that separates the data points. The choice of kernel function depends on the problem's nature and has a significant impact on the model's performance.

The key parameter of the SVC algorithm is C (Cost or Regularization parameter) which determines the trade-off between achieving a larger margin and minimizing classification errors. A high value of C implies low tolerance for misclassifications and results in a smaller margin. A lower value of C allows for more misclassifications while maximizing the margin, potentially yielding a smoother and less complex model (Cortes and Vapnik, 1995).

[1]Figure's reference: `https://medium.com/it-paragon/support-vector-machine-regression-cf65348b6345`

Model selection and validation in the context of SVC involve selecting the optimal value for the cost parameters to minimize validation errors and increase generalization ability. Techniques such as k-fold cross-validation and grid search can be employed to systematically explore a range of parameter values and evaluate their performance.

2.2.4 Support Vector Regression (SVR)

Support Vector Regression (SVR) is a popular machine learning algorithm used for regression tasks, based on the principles of SVMs. It aims to predict a continuous target value given input features. The main idea of SVR is to find a hyperplane that best approximates the underlying function of the data, while keeping the error within a certain margin (Drucker et al., 1996, Vapnik, 1999).

In SVR (Fig. 2.4(b)), instead of classifying the data points as done in SVMs, the algorithm tries to fit a function, $f(x)$, that best represents the given data points, while ensuring that the difference between the predicted value and the actual value is below a specific threshold ϵ (epsilon) (Smola and Schölkopf, 2004). The goal is to minimize the generalization error while maintaining a certain tolerance level for prediction errors.

The SVR algorithm employs a pairwise loss function, where it penalizes only the errors larger than ϵ. This means that errors within the ϵ margin are considered suitable and not penalized. This mechanism allows SVR to provide a robust and smooth prediction.

SVR can also use different kernel functions to map the input data into higher-dimensional spaces, allowing it to model non-linear relationships. The choice of the kernel function (linear or non-linear) depends on the nature of the problem, and it can significantly influence the performance of the SVR model.

A key parameters of the SVR algorithm is C (cost or regularization parameter) which determines the trade-off between model complexity and the degree of tolerance for errors outside the ϵ margin. A high value of C corresponds to a stricter penalty for errors, leading to a more complex model. A lower value of C implies higher tolerance, often resulting in a smoother and relatively simpler model.

To find the best SVR model for a problem, it is essential to perform model selection and validation using techniques such as k-fold cross-validation. By systematically determining the optimal values for parameters like C, ϵ, and kernel-specific parameters, we can generate a model that performs well on unseen data.

2.2.5 Transformer-based Natural Language Processing

Transformer-based techniques in NLP signify a noteworthy progression within the NLP domain. They are founded on the deep learning framework called the Transformer, which was originally introduced by Vaswani et al. (2017). Transformers

stand as neural network architectures explicitly crafted to manage sequential data, rendering them especially potent for NLP applications.

These models hinge on an innovative mechanism referred to as "attention," enabling the model to assign varying degrees of importance to different words within a sequence when formulating predictions. While the original Transformer structure encompasses both an encoder and decoder, it is common in NLP tasks to employ solely the encoder component, as it aligns well with tasks such as language modeling, text classification, and sequence-to-sequence operations (Vaswani et al., 2017). In the following, we provide a concise introduction to one of the pivotal transformer-based models, the BERT model which we use for task clarity assessment in Chapter 4.

BERT Models

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art transformer-based architecture for NLP tasks. Google unveiled this model in 2018, and since then, it has had a transformative impact on numerous NLP applications such as question answering, text classification, named entity recognition, and others (Devlin et al., 2019).

BERT model acquires context-aware word representations by utilizing vast quantities of unlabeled textual data. The unsupervised/self-supervised training process is based on a masked language modeling objective, in which the model learns to predict masked words within a given sentence. Additionally, BERT is trained using a next sentence prediction objective to comprehend the connections between successive sentences. BERT also tokenizes input text into subword units allowing the model to handle out-of-vocabulary words and capture more fine-grained information (Devlin et al., 2019).

Bert-base-uncased and Bert-base-cased are two variants of the BERT model, characterized by their treatment of text capitalization. Bert-base-uncased is a case-insensitive model version, converting all text to lowercase during both pre-training and fine-tuning phases. Consequently, words are treated as identical, regardless of their initial capitalization (e.g., "Apple" and ´´apple" are regarded as the same). In contrast, Bert-base-cased is a case-sensitive model that retains original word capitalization during pre-training and fine-tuning. As a result, the model can differentiate between words with distinct capitalization (e.g., "Apple" and "apple" are considered separate words) (Devlin et al., 2019). The choice between Bert-base-uncased and Bert-base-cased models relies on the particular task, text data characteristics, and dataset capitalization patterns. It is crucial to take these factors into account when choosing the suitable BERT variant for a specific NLP application.

2.3 Related Work

In this section, we offer a brief overview of the relevant literature concerning the use of NLP methods for assessing clarity and the solutions aimed at enhancing crowdsourcing procedures. Section 2.3.1 outlines the studies that evaluate the clarity of natural language text in diverse domains using NLP techniques, while section 2.3.2 leverages the summary of the related work as presented in the publication by Nouri et al. (2023). It offers a comprehensive outline of the research aimed at enhancing crowdsourcing workflows, models, and processes.

2.3.1 Clarity Assessment through NLP Techniques

Research efforts focused on assessing text clarity or quality using NLP methods revolve around the application of computational techniques to measure the understandability and readability of written text. Additionally, they involve the assessment of text quality concerning specific objectives or goals. These academic inquiries commonly make use of NLP algorithms and utilities to analyze and evaluate the lucidity of written materials (Schütze et al., 2008).

Skitalinskaya et al. (2021) introduced a novel approach to evaluating the quality of arguments by taking into account various revisions of the same claim. In particular, they collected over 300k pairs of claim revisions from the *kialo.com* platform, each representing an enhancement in quality. Using this dataset, they aimed to assess the quality of claim revisions and rank a set of revisions. To assess the quality of revisions, they employed traditional logistic regression models based on word embeddings, as well as transformer-based neural networks like BERT and SBERT. For the ranking task, they used the Bradley-Terry-Luce model and SVM^{rank}. Finally, they conducted a comprehensive error analysis for different revision types and varying numbers of revisions to determine the reliability of claim quality assessments.

Usmani et al. (2020) also proposed a system named "Clarity," an unsupervised, data-driven system designed for product assessment, that conducts automated and continuous analysis of the competitive landscape of products within a marketplace. The system employs sets of pre-trained Word2Vec models readily available off-the-shelf on a collection of online content to calculate competitive results.

Besides, Joung et al. (2018) employed text mining techniques to examine customer complaints and identify deficiencies in the company's products. In particular, they proposed a technique that integrates text mining with the Outcome-Driven Innovation (ODI) method in order to extract customer requirements from customer complaints.

Stab and Gurevych (2017) also introduced a method for assessing the quality of natural language arguments, emphasizing the sufficiency criterion. To overcome the challenges related to evaluating argument quality within argumentative writing

support systems, they developed feature-based SVMs and utilized Convolutional Neural Networks (CNN) to automate the identification of arguments that lack adequate support.

Persing and Ng (2013) employed binary SVCs to create computational models capable of assessing the thesis clarity of student essays. They also developed a system that can pinpoint the reasons behind the assigned score. They introduced five common errors that can hinder thesis clarity. The models' role is to detect which of these errors are present in a given essay. The evaluation of this clarity scoring model and the error identification system was conducted on a dataset comprising 830 essays, each annotated with both thesis clarity scores and identified errors.

In their study, Afful-Dadzie et al. (2014) conducted text analysis on user comments shared on social media platforms to conduct a comparative analysis of telecommunication providers in Ghana. On the other hand, Bhatt et al. (2015) monitored the overall sentiment trends over time for products. They achieved this by computing a sentiment score using user-generated content like reviews and comments.

2.3.2 Crowdsourcing Process Improvements

In the realm of crowdsourcing processes, the presence of suboptimal results submitted by crowd workers poses a persistent challenge in fully leveraging the potential of crowdsourcing (Weld et al., 2015). This challenge arises from a multitude of difficulties posed by all stakeholders participating in the process, including crowd workers, requesters, and the intermediary platform. Among these difficulties, the issue of unclear task design by requesters has been identified as a particularly significant factor (Manam and Quinn, 2018) and ambiguous task descriptions have yet been emphasized as a constant challenge (Khanna et al., 2010, Chandler et al., 2013, Gadiraju et al., 2017, Gaikwad et al., 2017, Wu and Quinn, 2017, Nouri et al., 2020).

The main focus of this thesis, therefore, centers around the development of a solution that addresses the challenge of unclear task descriptions written in natural language within the crowdsourcing work model.

The concept of text clarity has been explored in relation to readability and understandability, encompassing various aspects such as the syntax and semantics of text (Kincaid et al., 1975), problematic wording that causes semantic difficulties (Chall and Dale, 1995), and the use of statistical language models for assessment (Collins-Thompson and Callan, 2004). Computational analyses on text readability have also been conducted by Kevyn (2014).

In this section, we provide an overview of the models and workflows that have been developed to advance the state of the art in crowdsourcing marketplaces. The objective of our thesis is to explore the potential of enhancing task clarity in crowdsourcing through the use of an automated task clarity assessment approach. There-

fore, we outline the relevant literature that has designed workflows, methods, and tools with the same purpose.

Workflows

A workflow proposed by Salehi et al. (2017) addresses complex writing tasks by involving a series of interactions between workers and requesters. Initially, workers post their questions about the task, which are then discussed with the requester. Workers subsequently create a draft, which is rated and discussed by the requester. Based on the feedback, workers make edits and submit the final version. While this workflow allows for clarifying task instructions, it can be time-consuming and costly. It heavily relies on effective communication between workers and requesters, as well as a well-structured feedback mechanism, which can be challenging to achieve.

Similarly, TaskMate, introduced by Manam et al. (2019) offers a collaborative workflow for improving task description clarity. In this approach, workers identify ambiguities in the task description through questions and propose multiple answers for each question. Other workers then vote on the answers that are likely to clarify the ambiguities, leading to the creation of a clearer task description. TaskMate aims to reduce the burden on requesters in creating clear instructions. However, its effectiveness relies on the assumption of reliable worker collaboration and overall work quality.

Gaikwad et al. (2017) propose a workflow called Daemo, which allows requesters to post multiple instances of their tasks and gather feedback from workers. This feedback serves as a basis for optimizing the task description. Experimental results show the effectiveness of this approach in principle. However, it is important to consider that this method incurs costs in terms of time and money for the pilot step. Additionally, the approach may not be suitable for large crowds with diverse backgrounds and skills, as it relies on the subjective judgments of a limited number of workers providing their feedback.

On the contrary, our computational approach presented in Chapter 5 offers an automated tool that provides a faster and more cost-effective workflow for requesters. Unlike previous approaches, our proposed workflow completely eliminates the involvement of workers in the process. This independence from workers addresses various challenges discussed in prior work, such as workers' low-quality submissions, difficulties in establishing a good requester-worker relationship, and the absence of an effective feedback system. It is important to note that our approach can still be complemented by workflows or interventions that incorporate the opinions and feedback of workers if deemed necessary.

Models and Methods

Several researchers have proposed methods and models to define and analyze task clarity in crowdsourcing. For instance, Gadiraju et al. (2017) introduced a computational model that highlights predictive features like role clarity and goal clarity as key aspects of task clarity. They also demonstrated a way to measure clarity in the context of crowdsourcing tasks.

Wu and Quinn (2017) examined the impact of guidelines on workers' comprehension of task quality and its subsequent effect on task outcomes, including accuracy, trust, throughput, and worker fulfillment.

Papoutsaki et al. (2015) examined novice requesters in collecting data, how they design the task, what factors they consider, and alike. Their study contains valuable lessons for inexperienced requesters.

Khanna et al. (2010) indicated that the user interface, task descriptions, and the cultural background of workers challenge them with low digital capabilities from accepting and completing tasks on MTurk. They suggest simplifying the user interface and task descriptions, as well as language localization, to harness the full potential of workers.

Finnerty et al. (2013) showed that clear instruction and a simple task design that encourages workers' awareness enables higher quality results.

In order to develop computational methods, we expanded upon the clarity aspects identified by Gadiraju et al. (2017). We annotated the same dataset used in their study to explore the feasibility of assessing clarity flaws in task descriptions solely through the processing of plain text. We are confident that the assessment study we present in Chapter 4 will contribute to the advancement of automated writing assistance tools. The use of such assistance tools will not only aid inexperienced requesters in improving the clarity of their task descriptions but also assist workers in gaining a better understanding of the tasks, ultimately resulting in more precise outcomes.

Tools

Manam and Quinn (2018) presented WingIt for vague task instructions. The tool relies on the workers' comprehension and intuitions of the task conditions as well as the requesters' expected results. WingIt allows workers to connect to the requester and request clarifications through questions ("Q&A") with the best possible answer that resolves the ambiguity or to modify the descriptions ("Edit") directly. The communication is either synchronous, where the worker waits for an answer from the requester within three minutes, or asynchronous, where the worker submits the result supposing the requester approves the answer.

SPROUT (Bragg et al., 2018) is another tool that gathers confusing questions and employs recommendations from crowd workers to modify unclear parts of task

instructions. It supplies the requesters with queries and enables them to prioritize them. Such tools can support inexperienced workers at the cost of considerably extra time on the workers' side and additional money on the requester's side. However, the risks of misunderstandings and wrong interpretations of workers remain, which perhaps causes rejections and, consequently, a bad reputation.

Turkomatic (Kulkarni et al., 2011) offers a more worker-oriented technique based on a price-divide-solve algorithm. Turkomatic employs the crowd to decompose complicated tasks and complete the sub-tasks via step-by-step guidance. To succeed, it depends on knowledgeable workers, the requesters' leadership, and a high-quality feedback strategy.

Similarly, Chang et al. (2017) follows the idea of a collaborative system called Revolt, designed explicitly for image-labeling tasks with insufficient or ambiguous descriptions. Revolt allows multiple workers to label the task with the provided steps and access the instructions offered by other workers. In a conflict, workers revise the labels for the image based on other workers' instructions.

Alike, the system Fantasic (Gutheim and Hartmann, 2012) evaluates a task design to support novice requesters. It gathers task requirements from requesters to produce and display a task description before publishing it on the platform. However, this is restricted to a limited set of task types.

Other examples are Soylent which is a plug-in for Microsoft Word using which workers revise, shorten, and proofread documents by hiding the sophistication of task specification (Bernstein et al., 2010), along with CrowdForge (Kittur et al., 2011), and Crowd4u (Ikeda et al., 2016). The two latter aid in decomposing sophisticated instructions written in natural language into small sub-tasks for crowdsourcing platforms, though they lack generalizability to all task types.

Finally, TurKit supports requesters with task deployment iteratively on MTurk (Little et al., 2010) with the architecture designed to avoid acquiring redundant results by saving intermediate submissions. TurKit assumes that requesters properly determine the task decomposition logic in all cases.

In contrast to all discussed tools, in this thesis, we develop models using natural language processing methods to automatically score the unclarity level of task descriptions without involving workers and platforms in the process or relying on worker-requester communication quality. Employing these models, we introduce an automated tool that assists requesters in iteratively identifying and improving their task description clarity flaws before publicizing them on the platform. The tool is fully independent of workers' interaction, following a more efficient approach concerning workers' and requesters' time and money to obtain explicit task descriptions. Besides, this approach bypasses various challenges arising from the difficulties of requester-worker communication.

In the forthcoming chapters, we will present the approach we adopt to initially gain a comprehensive understanding of the crowdsourcing challenges and their prevalence in Chapter 3. We then analyze the findings from previous studies and define the focus of the thesis, namely unclear task description in natural language written by requesters. Consequently, in Chapter 4, we carry out a study to evaluate the effectiveness of natural language processing methods in addressing the issue. Building on the insights from the assessment study in the prior step, in Chapter 5, we proceed to develop the required models to create an automated interactive tool. To gauge the effectiveness of our solution, we assess our tool in relation to our primary objectives.

Chapter 3

Crowdsourcing Challenges

In this chapter, we present the methodology employed to examine the challenges associated with crowdsourcing, serving as a preliminary step to define the specific focus of this thesis. Our study encompasses a comprehensive exploration of both the theoretical perspective from existing literature and the practical viewpoint of crowd workers, enabling us to gain a comprehensive understanding of the challenges encountered by workers in crowdsourcing processes (detailed in Section 2.1). This investigation uncovers various areas in need of solutions to enhance the effectiveness of crowdsourcing. Subsequently, we narrow our focus to a crucial problem that serves as the central focus of this thesis. This chapter draws upon the research published in (Nouri et al., 2020) which elucidates the examination of crowdsourcing challenges.

Crowdsourcing harnesses the intelligence, skills, or information from an extensive collection of unknown individuals to produce creative ideas and innovative solutions. Crowdsourcing models aim to provide companies or individuals (i.e., requesters) with collaborative services and solutions for their problems. It also provides individuals (i.e., crowd workers) with remuneration, competence, or reputation in an anonymous setting with flexible time and workplace.

Crowdsourcing's promises interest a large number of requesters and crowd workers from diverse areas in academia and industry such as psychology, business, information systems, software development, and organization management. However, fulfilling such promises in an extensive interdisciplinary network of requesters and crowd workers with various knowledge backgrounds and diverse cultures has been challenging (Gupta et al., 2014). Moreover, the initial promises are influenced by the quality of the requester-worker relationship, their mutual trust, and satisfaction in their collaboration (O'Neill and Martin, 2013).

With a dependable Internet connection, crowd workers have the flexibility to engage in a wide range of tasks that suit their preferences, spanning from intricate computer programming, data analysis, and graphic design to more straightforward micro-tasks. This global workforce has the freedom to work from anywhere in the world. However, there are also some risks from committing to such work concern-

ing their employment status, satisfactory remuneration, social security, and other benefits (Bederson and Quinn, 2011). The chances and risks the workers face raise questions about their experiences of this form of work and what challenges they face in their daily work life.

Digital services (websites or apps) facilitating crowdsourcing business models, i.e., crowdsourcing platforms, offer the technical infrastructure necessary for requesters to publicize jobs to a massive pool of possible workers, retrieve and assess the results of completed tasks, and pay workers for the solutions they submit. Conversely, these platforms also provide crowd workers a central location to gather jobs from various requesters, a process to provide work products, and the financial and technological foundation to get paid for submissions.

Amazon Mechanical Turk (MTurk)[1] is the most well-known micro-task platform among micro-task platforms such as microWorkers[2], Crowd4U[3], Prolific[4], Upwork[5], iStockPhoto[6], and Figure Eight[7] (formerly CrowdFlower). Workers of MTurk (known as turkers) have created online discussion boards or communities, such as MTurk Crowd, TurkerView, TurkerNation, and Turkopticon, to discuss various aspects of work, such as tasks (known as Human Intelligence Task (HIT)), requesters, earnings, tools, and strategies.

Crowd workers, similar to requesters, face complex issues that they seek to discuss and resolve with the support of their communities. According to the shared experiences on discussion forums, both parties face challenges in practice that potentially harm their relationship. With the general objectives of the "Digital Future" research program in mind (introduced in Section 1.1), our primary aim is to gain a comprehensive understanding of the challenges that hinder the effectiveness of crowdsourcing processes. Leveraging this deep understanding and technological tools, we strive to enhance these processes and overcome the identified obstacles.

3.1 Approach

Aligned with the overarching goal of the "Digital Future" project, our task involved conducting an extensive examination of the challenges encountered in crowdsourcing processes. Given the significant role played by crowd workers, who form a majority of human participants in the process, and their crucial contribution in delivering high-quality solutions and products to requesters, we specifically directed our focus towards investigating challenges from the workers' perspective.

[1] MTurk's homepage link: `https://www.mturk.com`

[2] microWorkers' homepage link: `https://www.microworkers.com`

[3] Crowd4U's homepage link: `https://crowd4u.org/en/`

[4] Prolific's homepage link: `https://www.prolific.co`

[5] Upwork's homepage link: `https://www.upwork.com`

[6] iStockPhoto's homepage link: `https://www.istockphoto.com/de`

[7] Figure Eight's homepage link: `https://visit.figure-eight.com/People-Powered-Data-Enrichment_T`

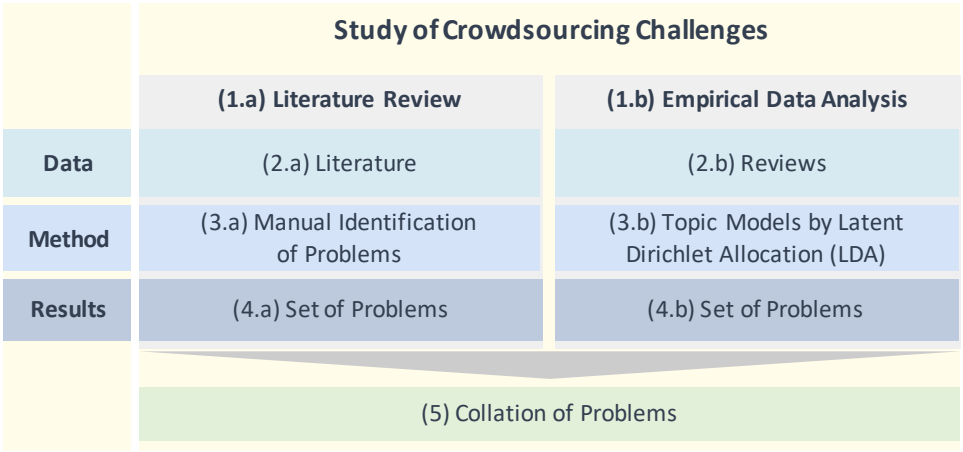


FIGURE 3.1: Overview of the approach adopted to study crowdsourcing challenges.

We followed a twofold approach to study challenges in crowdsourcing processes (illustrated in Fig. 3.1). We first conducted a literature review (Fig. 3.1(1.a)) to study the literature (Fig. 3.1(2.a)) and manually identify the challenges found by researchers (Fig. 3.1(3.a)), mainly from the theoretical point of view, which resulted in a set of discussed problems in the literature (Fig. 3.1(4.a)). Furthermore, we conducted an empirical data analysis (Fig. 3.1(1.b)). In this process, we gathered workers' reviews (Fig. 3.1(2.b)) from a discussion forum where crowd workers shared their concerns regarding difficulties, confusion, and unfair treatment experienced during their work in crowdsourcing tasks, and then mined discussed challenges among workers using topic models (Fig. 3.1(3.b)). Our hypothesis suggests that the discussions taking place in forums likely contain workers' problems and their practical significance that has not yet been recognized in the existing literature. Through empirical analysis, we identified a set of problems that have been extensively discussed among workers (Fig. 3.1(4.b)). We then combined the findings from the two studies (Fig. 3.1(5)) to address the two following research questions:

> **(RQ1)** What problems do workers face in the different phases of crowdsourcing processes?

> **(RQ2)** Which of these problems are most dominant in the literature and in the data respectively?

In the following, we explain the literature review and the data analysis step in detail. Next, we aggregate the findings from both steps and interpret them in detail.

3.2 Literature Review

In order to compile the challenges discussed in research publications, we conducted a thorough literature survey to examine studies that specifically addressed the problems encountered by crowd workers in various steps of crowdsourcing processes. The data for this survey was gathered primarily through studies, supplemented by interviews conducted with crowd workers. In this section, we outline the methodology used for selecting relevant articles, provide an overview of the articles reviewed, and present the findings obtained from this analysis.

Literature

For article selection, we employed the Google Scholar search engine to perform searches for articles with titles containing relevant keywords such as "crowd workers," "crowdsourcing," "crowdsourcing platforms," along with terms like "difficulties," "issues," "problems," and "relationships". A total of 102 articles were gathered through the initial selection.

From this pool of candidates, we carefully reviewed the abstracts, introductions, and conclusions of each article to identify those that specifically addressed the challenges faced by crowd workers in crowdsourcing processes. Furthermore, we employed an iterative process where we examined the references of the selected articles to identify additional potential candidates. As a result of this process, we identified 33 articles that explored crowdsourcing challenges from both theoretical and practical perspectives. The practical viewpoint was explored through various methods such as questionnaires, interviews, or surveys conducted on crowdsourcing platforms. In line with the data used for our empirical data analysis, most selected studies have been conducted on MTurk which is recognized as the largest and most renowned micro-task crowdsourcing platform (O'Neill and Martin, 2013).

The 69 excluded publications (detailed in Appendix A, Section A.1.1) can be grouped according to their field of study: (a) information systems and human computation, including research on crowdsourcing processes, models, methodologies, the risks and benefits of their applications in other domains, and quality management models for platforms and involved stakeholders, (b) psychology, including studies on behavioral and job attitudes, and (c) business and organization management, including studies on business values, and workforce management. Few other articles were also from the law and sociology area (Nouri et al., 2020).

Method

Altogether, we reviewed 33 articles discussing challenges in crowdsourcing processes. We manually collected the problems mentioned in the selected articles

TABLE 3.1: The 14 problems found through the literature review were categorized based on their respective causes.

Category	Problem
Task Design	1) Ambiguous Instructions, 2) Malfunctioning Environment, 3) Workload Misestimation, 4) Low Payment, 5) Privacy Violation
Task Operation	6) Missing Responses, 7) Minor Feedback, 8) Mean Comments
Task Evaluation	9) Unfair Rejections, 10) Late Payment, 11) Unjustified Blocking
Platform	12) Poor Mediation, 13) Imbalance Power, 14) Poor Tooling

and classified the found problems into the three phases (described in detail in Section 2.1) of the crowdsourcing process: Task Design, Task Operation, and Task Evaluation. The class of the problems shows the stage when they have been created, not necessarily when the workers experience them. The challenges associated with crowdsourcing platforms are also packed in one group. To find the significance level of found problems, we counted the occurrence of each problem. We could then compare their significance with those identified from the data analysis approach.

Results

As published in (Nouri et al., 2020), Table 3.1 shows 14 problems that were identified in the literature review step. Eleven problems are classified in the three main phases of crowdsourcing processes such as task design, task operation, and task evaluation. The remaining challenges are platform-related problems, thus grouped in another category. In the following, the three phases and their related problems are listed with numbers increasing from 1 to 14 in parentheses (Nouri et al., 2020).

Task Design Problems

Here, we discuss the problems that can emerge in the task design stage of the crowdsourcing process when those making requests fail to adequately design their tasks, leading to issues with the underlying framework, unfair assessments, and the essential information required for task completion. For each problem, we present

the number of articles that discuss it based on our literature review, along with detailed information derived from multiple examples within those articles.

(1) Ambiguous Instructions Five out of 33 reviewed literature report that workers deliver low-quality results due to unclear task descriptions, where it is difficult to understand what the task is about and what submission the requesters desire.

Fowler Jr (1992) investigated the effect of unclear terms on the quality of survey data. They reported that although survey questions should be easy to understand, there are no established protocols to ensure that each important term is consistently understood. They showed that clarifications on the key terms' definitions resulted in very different comprehension and, consequently, final results by workers. The implication is that estimations are likely biased when terms are unclear. Finally, they suggested that one strategy to lessen systematic error in survey results is to evaluate survey questions to identify important terms that are not consistently understood and to define unclear terms.

Chandler et al. (2013) also discussed the risks and rewards of using online markets to facilitate crowdsourced human computation. They talked about the benefits and drawbacks of these marketplaces, focusing in particular on the accuracy of the crowdsourced information gathered through MTurk. They related the reliability and the quality of submitted results to poorly communicated instructions, lack of workers' comprehension of descriptions, and requesters' neglect of the effect of the clear instructions on the final results. Besides, Khanna et al. (2010) examined the usability constraints restricting users with little knowledge from completing simple tasks on Mechanical Turk. They showed that innovative design components, such as improved user interfaces, simplified and clear task instructions, and language localization, are crucial in crowdsourcing marketplaces.

Similarly, Gaikwad et al. (2017) emphasized that ambiguous instructions frequently leave workers and requesters unable to trust one another's quality, and this causes misalignment of their mental models of tasks and failure of crowdsourcing processes. Gadiraju et al. (2017) highlighted task clarity as an essential characteristic of crowdsourcing tasks. They discussed that poor task instructions directly impact workers' poor comprehension of the work and deliver low-quality results, while requesters are unaware of problems with their task design. Thus they interpret unsatisfactory work as proof of malice and withholding rewards.

(2) Malfunctioning Environments Six articles reported that the workers become frustrated from facing technical errors in the task interface created by the requesters. This problem occurs when workers either use the interface to do the task or submit their submissions. Both lead to unsuccessful submission and useless effort (Silberman et al., 2010b, Silberman, 2015).

Bederson and Quinn (2011) provided a summary of the ethical and observed issues surrounding online labor as well as a set of recommendations that enable more positive relationships between workers and requesters. They specifically reported that the workers had faced many tasks that wasted much time navigating task interfaces, waiting for network activity, or resolving technical issues, the cost of which is carried by the workers. Brawley and Pury (2016) also applied Industrial/Organizational (I/O) psychology principles to study MTurk worker job satisfaction, information sharing, and turnover. They also reported that the best and worst requester behaviors (e.g., relationship building, unfair pay) affect worker satisfaction. They mentioned that workers face technical difficulties causing rejections as one of the negative requester behaviors.

Moreover, McInnis et al. (2016) focused on how turkers governed the hazards of rejected work resulting from task clarity flaws, poor design, and implementation. They reported that workers are frustrated about the consequences of technical errors in task interfaces or Application Programming Interfaces (APIs) when performing the tasks. Berg et al. (2018) also published one of the first research studies on the conditions of work on micro-task platforms that covered many facets of crowd work. The research outlined the advantages and disadvantages of crowdsourcing and proposed some guidelines for improving working conditions on platforms for digital labor. They discuss that workers should not be responsible for the costs associated with lost time or work caused by technological issues with tasks or platforms, yet the responsible party should compensate the employee fairly.

(3) Workload Misestimation According to three studies, crowd workers also complain that requesters wrongly estimate the required effort and time to finish the task which leads to timeout and ineffective attempts to complete the tasks (Silberman, 2015).

Silberman et al. (2010a) explored fundamental questions about the workers, their motives, and their relationship to the requesters who pay them as professional human computation power. They also discussed the problems and unfairness that workers face at work. For instance, requesters have a poor understanding of the task and post a task with an unreasonable short deadline, which ends before workers can finish it. Thus, workers do not receive payments for the time they spend, and their reputation statistics are impacted rather than the requesters'. Gupta et al. (2014) studied the crowd workers in India, i.e., who they are and how they maintain their work-life balance in crowdsourcing marketplaces. They emphasize that the workers complain about the tasks that take a long to complete for inappropriately low compensation.

(4) Low Payment Four studies discussed that the tasks with low hourly payment ratios may be created unintentionally due to the requesters' misestimation of ef-

fort and time or intentionally due to their unwillingness to pay fairly for the task. Eventually, it may affect the workers' motivation and satisfaction (Ross et al., 2010, Gadiraju et al., 2017).

Silberman (2015) studied crowd works of MTurk platform and discussed various aspects of crowd work challenge and their impact on future work. This study extensively reported that tasks with a low average hourly wage on MTurk are accepted by the workers who have MTurk as the only source of income to pay for their basic needs. Berg et al. (2018) also discussed that the payment on crowdsourcing platforms, including MTurk are quite low. They stated that in 2017, the average hourly wage for a worker was USD 4.43 when only paid work was taken into account. However, when both paid and unpaid hours were included, the average wage was USD 3.29 per hour.

(5) Privacy Violations Ten literature also report that tasks violate workers' privacy by asking, processing, and misusing workers' personal information, such as their real name, worker ID, email address, etc., in various ways (Kang et al., 2014, Silberman, 2015, Edlund et al., 2017, Shu et al., 2018). Such attempts even break MTurk's term of service (TOS).

Lease et al. (2013) discussed how the MTurk system design reveals a surprising amount of data about many MTurk Workers, some of which may be Personally Identifying Information (PII), and assessed the potential multi-faceted impact of such PII exposure for workers, requesters, and MTurk. In this study, workers mentioned that their worker ID is associated with their Amazon product reviews or wish lists. If workers perform poorly, a disappointed requester may use their PII to damage their reputation in other online communities or even attempt to identify the person's actual location. Kang et al. (2014) and Halder (2014) investigated the potential issues with data protection, privacy, and security in crowdsourcing platforms. They reported that the workers' personal information and data privacy are not properly protected on the crowdsourcing platforms.

Vakharia and Lease (2015) also conducted a qualitative content analysis of seven alternative platforms, including MTurk and outlined the main problem categories with MTurk versus platform features from the content analysis. The study showed that workers disclose their personal information to reach 'trusted member' status or to voluntarily let requesters track their cross-platform reputation. Besides, Durward et al. (2016) investigated four main aspects of crowdsourcing processes, such as privacy, accuracy, property and accessibility and discussed different dimensions of privacy including personal information, surveillance, communication. They mentioned that the crowdsourcing platforms gather a huge amount of data from the crowd workers. They then use, store, and analyze all the data to offer the requesters the best solutions to their issues.

Sannon and Cosley (2019) studied why crowd workers decide to provide their personal information and how they navigate the risks of personal information exposure. Several privacy concerns and infractions have been mentioned by turkers, including data gathering and profiling, unauthorized data usage, intrusive stalking and spamming, and misleading tactics like phishing and frauds. Xia et al. (2017) provided the findings of an online survey on privacy concerns, including data aggregation, profiling, and scams, as well as privacy expectations on MTurk and personal privacy losses from phishing, malware, stalking, and targeted marketing (e.g., screening tasks). The fact that respondents from different nations and regions shared similar experiences with privacy difficulties raises the possibility that these vulnerabilities may trouble the entire MTurk platform.

Task Operation Problems

Similarly, we here delve into the problems that arise in the task operation phase of the crowdsourcing process, where workers engage with tasks and may require feedback or responses if they encounter difficulties with the task or have inquiries. For each problem, we will provide the number of articles discussing it based on our literature review, along with detailed information derived from various examples within those articles.

(6) Missing Responses 13 literature also show that requesters often are careless in responding to task-related or solution-related inquiries from workers that may cause low-quality or invalid submissions by workers (Silberman, 2010, Chandler et al., 2013, Gupta et al., 2014, Alagarai Sampath et al., 2014, Silberman, 2015, Brawley and Pury, 2016, Deng and Joshi, 2016, Schwartz, 2018).

Silberman et al. (2010a) mentioned in their report that the requester often ignores the workers and rejects their submissions when their task is poorly designed and its instruction is unclear. Silberman et al. (2010b) conducted surveys to find crowdsourcing problems and then outlined some projects to address the problems. In the report, they stated that some of the responses to the survey on crowdsourcing challenges expressed workers' frustration due to unresponsive requesters to their emails, mainly inquiring about reasons for unfair rejections. Bederson and Quinn (2011) also emphasized that workers complain about unresponsive requesters to emails (which are mediated by MTurk to maintain anonymity) when technical errors arises.

Dow et al. (2012) examined whether a task-specific feedback system for crowdsourced work assisted crowd workers in learning, perseverance, and producing better results. They stressed that the majority of current micro-task platforms support asynchronous feedback, often with a several-day delay. However, at that point, workers might be more concerned with getting paid than improving previously delivered work. Berg et al. (2018) stated that there are often little or no

possibilities for communication between platform management, workers, and requesters. Workers are theoretically able to contact the platform management or the requesters, but in practice, it might be difficult to obtain the right contact details, and responses may be delayed, unsatisfactory, or missing.

(7) Minor Feedback According to four articles, crowd workers sometimes ask for requesters' feedback on their results to increase the fulfillment level of quality results' requirements and, consequently, the chance of acceptance. However, requesters often give either no or little feedback to submitted results. Berg et al. (2018) discussed the lack of proper feedback in the workflow, which helps workers to improve their results and increase the chance of correcting their mistakes and consequently the approval and payment. Besides, Dow et al. (2012) comprehensively discussed the lack of proper feedback mechanisms in crowdsourcing processes and emphasized the necessity of a well-structured synchronous or asynchronous feedback system in order for workers to receive expert feedback on their work, which yields better overall outcomes and helps workers to improve their performance over time.

Gaikwad et al. (2017) introduced Daemo's Boomerang reputation system, aimed to increase the likelihood that workers and requesters work together on a solid mutual trust in the future. In this work, the lack of well-designed feedback is outlined, and the importance of the feedback system on improving quality results and building a good relationship between requesters and workers is highlighted. Schwartz (2018) also discussed the limited communication and and lack of a constructive response to work-related inquiries from workers.

(8) Mean Comments Four studies also report that some responsive requesters communicate unfriendly with workers or give mean comments on the questions or the submitted results. Berg et al. (2018) highlighted the unfriendly behaviour of malicious requesters with leaving unfair and mean comments on workers' submissions. Martin et al. (2014) analyzed the content of a forum for MTurk users with a focus on the workers' judgment of tasks' values and the influenced factors for accepting and completing a task. They reported that workers who mentioned the demeaning comments also noted such challenges are extra damages added to other difficulties they face in crowd work.

Xia et al. (2017) discussed the requester-worker confrontation challenges, some of which led to dramatic fights sometimes because of a rejection. One worker told the story of a fight where eventually, the requester found the worker's social media account using his email address and left harassing comments for the worker. Brawley and Pury (2016) revealed the cases where the workers experienced receiving unnecessary rude comments on rejected results. The workers mentioned

they could accept the rejection of unsatisfactory results but not the mean behavior from the requesters.

Task Evaluation Problems

Similar to previous categories, we explore the problems that surface in the task evaluation stage of the crowdsourcing process. These issues stem from requesters who do not fairly evaluate the work submitted by workers and do not provide adequate explanations for rejections, non-payment, or delayed payment of tasks. For each problem, we present the number of articles discussing it based on our literature review, along with comprehensive information derived from multiple examples within those articles.

(9) Unfair Rejections 13 articles discuss this problem in two ways. On the one hand, 11 of them report that workers receive (a) *harsh evaluations* carried out on their submissions that have been created to the best of the workers' capabilities either by requesters or by automatic algorithms (Porter, 2004, Bederson and Quinn, 2011, Peng et al., 2014, Guth and Brabham, 2017). On the other hand, 10 articles mention that there are often (b) *missing explanations* for the rejections provided by requesters. Crowd workers receive no or vague justifications for the evaluation they achieve, which makes them finally disappointed (Porter, 2004, Peng et al., 2014, Martin et al., 2014, Silberman, 2015, Guth and Brabham, 2017).

Silberman et al. (2010a) broadly covered the unfair rejection problem and stated that the possibility of paying workers after evaluating their results makes crowdsourcing highly attractive to requesters, while workers are left vulnerable to the whims of requesters or, more often, the employers' evaluation software when determining the quality of their assignments. Requesters find it impractical to manually review submissions due to the work volume. They often confuse rejected workers with generic statements outlining reasons for rejection since they hire hundreds or more workers simultaneously. In the worst-case scenario, pernicious requesters will post numerous high-paying assignments, obtain the work, and then reject it to get free work, which decreases the workers' acceptance rate and their effective wages and make them feel vulnerable.

Silberman et al. (2010b) reported according to a worker, the annoyance of rejected work for no good reason is one of the biggest concerns in crowd work. They also emphasized that the workers believed requesters arbitrarily reject the assignments and often without providing a reasonable argument. Bederson and Quinn (2011) also stated that the unjustifiable and arbitrary rejection of good work was underlined as one of the biggest causes of workers' frustration. They discussed the lack of a rebuttal process and that workers could avoid working for that requester in the future. In this setting, the destruction of trust between requesters and workers is inevitable.

Irani and Silberman (2013) introduced Turkopticon, a system that enables workers to review and publicize their relations with requesters. Turkopticon serves as a shared infrastructure that encourages workers' cooperation and allows them to write and utilize reviews of requesters when choosing their tasks on MTurk. Irani and Silberman (2013) argued that the requester could reject or pay for it after a worker submits his assignment. MTurk's participation agreement guarantees requesters full intellectual property rights over submissions regardless of acceptance. Requesters verify worker submissions using automated methods according to a known response or the majority rule and automatically reject the other submissions irrespective of correctness. The full intellectual property rights agreement enables wage theft by unfair or careless requesters who reject and use the work without workers having little legal recourse against that.

Gupta et al. (2014) stated that unfair rejections harm workers' reputations on platforms and, consequently, their approval rate, which is one of the workers' main concerns in crowdsourcing. It also leads to a decrease in the availability of good and highly-paid tasks to crowd workers, many of whom work full-time to support their basic needs in life.

Brawley and Pury (2016) characterized unfair rejections as applying majority rules for rejection decision, mass rejection and taking advantage of workers, rejection for working too quickly, rejections caused by technical problems and similar and introduced them as one of the main factors of workers' frustration and dissatisfaction in crowdsourcing processes. Berg et al. (2018) discussed that crowd workers often complain about their work being arbitrarily rejected and not paid, probably by dishonest requesters. A notable characteristic of micro-task platforms is the sheer volume of submissions and the tendency for automatic evaluation and, eventually, payment by algorithms rather than humans. This algorithmic evaluation process potentially increases instances of unfair assessment. For instance, the method may automatically reject the work of the one response that is different, even if it is correct when three workers complete a task and one of their results differs from the others.

(10) Late Payment According to three articles, crowd workers are unaware of how long they must wait to obtain the payment after submission. In some cases, the delay is long without workers knowing about the payment time (Bederson and Quinn, 2011, Silberman, 2015, Brawley and Pury, 2016).

(11) Unjustified Blocking Six studies indicate that requesters sometimes block workers without a clear reason or if they ask for explanations of why their results are rejected or information about when they will be paid (Gupta et al., 2014, Martin et al., 2014, Peng et al., 2014, Guth and Brabham, 2017). Brawley and Pury (2016) mentioned the long delay in payment and blocking workers for no reason as known

examples of the negative behavior of requesters based on the workers' judgment. Irani and Silberman (2013) also reported that workers demand faster pay from requesters while they can evaluate and pay workers within 30 days.

Platform Problems

Here, we discuss the problems related to the process management and platform's policy, specifically focusing on the role of platforms in mediating conflicts between requesters and workers. Similar to previous categories, we provide the number of articles discussing each problem.

(12) Poor Mediation According to six studies, workers can report challenges concerning task design or communication with requesters to the platform. However, it is unclear how the platform processes, and impossible to follow how it solves them. Workers believe that platforms poorly mediate the cases of unfair rejections, disagreements on payments, or complaints from workers against requesters (Silberman et al., 2010b, Khanna et al., 2010, Irani and Silberman, 2013, Vakharia and Lease, 2015, Silberman, 2015, Schwartz, 2018).

(13) Imbalanced Power One article mention that requesters and workers have unequal powers or rights. On the one hand, requesters can reject results and not pay the workers. On the other hand, there is no support or control for workers to change that; instead, they must bear a lower reputation (Irani and Silberman, 2013).

(14) Poor Tooling Besides, five studies discuss that platforms provide poor functionalities, and workers have no access to requesters' contact information (Chandler et al., 2013, Berg et al., 2018), and there are no automated tools and helpful quality control on workers' reputation, task payments, and term-of-service violations (Vakharia and Lease, 2015, Silberman, 2015). Workers are also not anonymous because their Amazon account and MTurk ID are linked and searchable on search engines (Halder, 2014).

3.3 Empirical Data Analysis

Apart from conducting a literature review to examine the problems discussed by researchers in crowdsourcing processes from a theoretical standpoint, we also employ a topic modeling method (explained in Section 2.2.1) to identify the challenges faced by workers in practice on a large scale. Specifically, we analyze workers' narratives and accounts shared on an online discussion forum, where they expressed their experiences and grievances related to their interactions with requesters. While the empirical analysis of the narratives offers valuable insights

into the challenges discussed among workers and how they may evolve over time, for the purpose of this study, we have chosen to focus solely on the problems that workers encountered and shared with their fellow crowd workers during the process and set aside the examination of temporal changes. This empirical analysis allows us to gain valuable insights into the real-life challenges encountered by workers in their day-to-day tasks.

In this section, we present the corpus that was utilized for our analysis, which not only serves the purpose of this study but also holds potential value for future research within the community. Additionally, we outline the methodology employed and share the findings derived from our empirical study conducted on the reviews provided by the workers.

Data

Regarding the data analysis, one potential source of data could have been the records of actual communication between requesters and workers, including emails and chat protocols. However, due to the unavailability of representative data of this nature and the need to respect privacy concerns, we opted to explore publicly available data instead. This public data may contain valuable information about the relationship dynamics between requesters and workers, as well as the challenges that workers encounter when collaborating with requesters.

Since most crowdsourcing platforms lack to provide workers with the opportunity to share their experiences with requesters, worker community forums have been built to enable workers to support each other and avoid facing unfair situations with requesters. Workers of MTurk were the pioneer in creating such forums, which provide the public data containing the desired information for our study. The well-known examples of such forums are TurkerView, MTurkCrowd, TurkerHub, TurkerNation (now on Reddit), and Turkopticon. For our analysis, we select the latter because its initial concept is to offer MTurk workers a place to talk about the challenges they face in the process and to support their rights in their relationship with requesters (Irani and Silberman, 2013).

Turkopticon works as a reputation system that provides an environment where workers write so-called reviews about the requesters they have worked with. In the reviews, workers tell the story of their experiences, both positive and negative, working with specific requesters on MTurk, referencing the requester's unique ID and the specific task they completed. These experiences encompass a range of information, including details about the tasks themselves (i.e., the environment, the workload, etc.), payments (e.g., fairness and delays), and evaluation results (e.g., rejections, communication details). Finally, workers provide ratings for the requesters and indicate whether they would recommend or not recommend them to other workers (two examples shown in Table 3.2). If provided, the other workers

Example Reviews from Turkopticon
recommended
"I was having problems with 2nd page of the HIT; no matter what 3 choices I selected for a category, I could not submit the HIT and the HIT eventually timed out on me. I messaged the requester explaining the problem and haven't received a response as of yet. However, now I am having no problems with the second page so perhaps the requester fixed the problem I was having."
not recommended
"I took a chance and only (thankfully) did 2 of his HITS to try them out. I did them EXACTLY as he stated that he wanted....I came back an hour or so later and had two rejections from him. Stating "Please, select the under eye area more precisely and follow the instructions above" I went to send him a message and this is what came up "Failed to send comments We're sorry, but this Requester could not be reached." "

TABLE 3.2: Example of *"recommended"* and *"not recommended"* reviews written by crowd workers while doing tasks on Turkopticon platform. Due to the data privacy, we have removed specific task details (e.g., task's title and the requester's ID o MTurk. Note: Tasks on MTurk platform are called HITs which stands for Human Intelligence Tasks.

can gain insights into the requester's reputation for fairness and the nature of the tasks involved and then make informed decisions about whether or not to engage in a particular task.

To create the data set, we crawled all 27,041 existing reviews on Turkopticon from February 2017 to November 2018 (the time of crawling). The data included reviews with the tag "recommended," "not recommended," and none. For the corpus, we collected all 8,610 reviews with the tag "not recommended" for this analysis (examples in Appendix A, Section A.1.2), as we hypothesized them to have information about problems that workers were facing and, thus, beneficial for the present study. Altogether, the average length of reviews counted 57.2 tokens, and the resulting corpus comprises 492,713 tokens and 15,921 unique words.

Method

In the data analysis, we aim to mine crowdsourcing problems from the review. For this purpose, we deploy standard topic modeling using Latent Dirichlet Allocation

(LDA), as detailed in Section 2.2.1. LDA generally uncovers the hidden, conceivably overlapping k topics in a huge collection of documents (Blei et al., 2003). Employing LDA, a cluster of similar documents (here: reviews) is about a topic modeled as a list of representative words in the cluster. In this study, each topic represents the problem that workers complain about in the community.

In general, it is necessary to clean noisy text data before utilizing it as input for a machine learning model. Text cleaning or pre-processing is an essential step that must be undertaken before applying NLP techniques to a corpus. Real-world texts authored by humans often include various elements such as special symbols, punctuation, specific grammar-related function words, and more. Thus, we first removed unnecessary data from the corpus before applying LDA by performing the seven following steps of pre-processing:

(a) Tokenization: The manner of splitting text into units called tokens is tokenization that can be of three types: a) words, b) characters, or c) sub-words (n-gram characters). In this step, we break down each review into individual words and other tokens.

(b) Removal of numbers: The task-specific reviews contain numbers like dates, times, prices, and similar. Such numbers do not convey particular information regarding the main problems workers point out in their stories. Thus, we remove such numbers from the corpus.

(c) Removal of stopwords: For instance, “at”, “to”, “which”, “you”, “is”, and “myself” are the function words that appear typically across all the documents in the corpus. Removal of such words that are not biased and, therefore, have no significance in the analysis is a typical pre-processing step.

(d) Removal of punctuation: Punctuation (e.g., “,”, “.”, “!”, “;”, etc.) are useless marks and symbols in the documents we discarded from the corpus.

(e) Lemmatization: Another common text-cleaning technique used in machine learning is lemmatization. In the process, we replaced the different inflections of the same word with its root word, called a lemma.

(f) Removal of the tag: The “not-recommended” tag that occurred in all chosen reviews in the corpus without adding helpful information was also removed.

(g) Removal of low-frequency words: Considering how LDA clusters documents in a vast set, the words that occurred only once in all 8,610 analyzed reviews were useless for our analysis, therefore, discarded from the corpus.

The corpus size was decreased to 179,539 tokens after pre-processing (3808 unique words). Finally, an input instance to LDA averaged 20.9 tokens.

The parameters of LDA, such as the number of topics k and the priors for the per-document topic distribution α and the per-topic word distribution η (explained in Section 2.2.1), have a significant impact on the algorithm’s performance. Following (Wallach et al., 2009a) and (Syed and Spruit, 2018), we set the parameters of LDA to a specific value to narrow down the problem of finding the best topic

model to the choice of best k, which is a challenge mainly due to the lack of knowledge about the number of problems (i.e., topics) existing in this particular corpus. Hence, we adopt an empirical approach to determine the optimal value of k. Employing LDA, we build 14 models with topic number $k \in \{2, \ldots, 15\}$ each of which execute 500 iterations repeatedly. To find the best k among the generated models, we compare the suitability of existing evaluation metrics.

As clarified in Section 2.2.1, we choose to focus on topic coherence metrics among the various topic model evaluation metrics—such as *harmonic mean* (Griffiths and Steyvers, 2004), *pairwise cosine distance* (Cao et al., 2009), and *KL divergence* (Arun et al., 2010) for determining the best k for a given corpus. This decision is grounded in our aim to produce easily interpretable topics related to workers' problems. Topic coherence metrics have proven effective in assessing topic model quality and are commonly applied in natural language processing and information retrieval (IR) research.

Given the significance of accurately interpreting the topics, particularly pertaining to workers' problems, we evaluate the 14 models we generated using three metrics that gauge topic coherence based on human judgment, as they are most relevant in this context. Concretely, we assess our models utilizing the following metrics:

1. The *Mimno metric* is established on the notion that a successful topic model should produce cohesive, understandable topics that can distinguish between various corpus articles. By comparing the co-occurrence of keywords within each topic to the co-occurrence of words across the entire corpus, the Mimno metric evaluates the coherence and interpretability of a topic model. This metric uses top-word co-occurrence statistics and functions in three main steps: (a) identifying various classes of low-quality topics, b) identifying explicit semantic issues in topic models, and (c) optimizing of the topic coherence (Mimno et al., 2011).
2. *Normalized point-wise mutual information (NPMI)*, is a measure to evaluate the mutual dependence between words in a topic and is used to estimate the coherence and interpretability of a topic. NPMI calculates the probability difference of word co-occurrences to their expectation by calculating the average NPMI score for all pairs of words within the topic. A topic that is cohesive and interpretable has a high average NPMI score, whereas, a low score suggests a less coherent or interpretable topic (Bouma, 2009).
3. The *hybrid coherence measure* C_v, which is a metric that integrates multiple measures of coherence to evaluate the interpretability of a topic by comparing the co-occurrence of words within the topic to some external reference. Hybrid coherence measures combine multiple coherence measures to take advantage of each measure's strengths and lessen the influence of any individual measure's limitations (Röder et al., 2015).

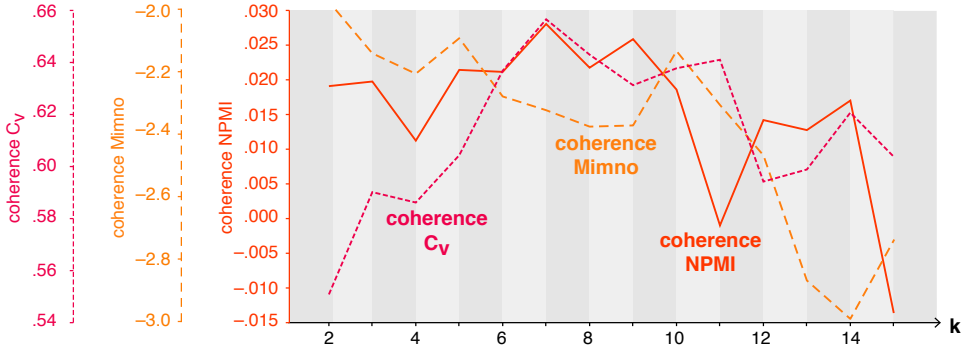


FIGURE 3.2: Topic coherence degree of our topic modeling approach according to the three applied evaluation metrics (Mimno, NPMI, and C_v) for different number of topics $k \in \{2, \ldots, 15\}$. Despite their different scales, the interpolated curves are shown in one plot, to make it easy to see the best k (Nouri et al., 2020).

We select k based on the metrics and then manually interpret each subject that resulted, as detailed in the following section.

Results

According to Figure 3.2 indicating the result of topic model evaluation, two of the three topic coherence metrics (NPMI and C_v) suggest $k = 7$, while Mimno votes for $k = 2$ and $k = 5$. We manually reviewed the assigned word sets to the topics for $k = \{2, 5, 7\}$ and then found the topic model of $k = 7$ the most well-separated and meaningful topics.

We first interpreted the seven discovered topics based on the 15 highest-probability words in the associated word lists, which can be in Table 3.3. To increase reliability, two authors of this work independently, one of whom did not know the literature review results. Both then joined to compose a problem label for each topic. We also examined the 20 highest-probability reviews for each topic to increase confidence in an accurate interpretation. In one case, this resulted in a minor label change. As for the literature review, we categorize the results of our data analysis based on the stages of the crowdsourcing process as follows:

Task Design Problems

The topic model we employed uncovered a collection of topics about shared experiences by workers regarding issues which primarily stemmed from inadequate requester performance during the task design phase of the crowdsourcing process. These workers advised others not to engage with the tasks offered by those requesters. Primary complaints included broken survey links, flawed processes for obtaining necessary completion codes for successful submissions, unfair tasks, and

TABLE 3.3: The seven crowdsourcing problems derived from topic modeling, along with the top 30 words of each topic.

#	Problem	Top 35 Words Obtained from Topic Modeling
1	Malfunction: Implementation errors	break, link, page, one, requester, survey, return, question, amazon, problem, report, get, dead, issue, try, answer, first, setup, accept, screener, review, set, check, work, would, click, open, message, already, see
2	Malfunction: Failed completion	code, survey, break, time, submit, complete, end, timer, get, completion, minutes, waste, error, finish, take, return, page, requester, short, say, given, message, provide, expire, response, give, work, minute, button, try
3	Workload: Bad time estimation	pay, minutes, time, low, long, take, question, one, bubble, page, way, would, even, end, get, bonus, answer, much, going, cent, minute, writing, survey, return, task, could, video, image, first, make
4	Workload: Too high work effort	writing, write, much, cent, worth, require, penny, paragraph, back, prompt, work, photo, want, single, throw, box, nope, date, every, warning, formatting, describe, clear, detail, story, bug, page, first, mislead, unclear
5	Violation of terms of service	tos, require, email, violation, information, ask, site, inquisit, firefox, address, error, personal, file, name, website, app, account, download, upload, collect, URL, google, info, scam, want, internet, explorer, another, phone, please
6	Underpaid tasks	underpay, pay, writing, screener, bad, screen, unpaid, avoid, requester, research, word, question, per, study, number, qualify, cheap, survey, may, contact, min, cent, many, participant, return, words, horrible, want, rate, ask
7	Unfair rejection	reject, requester, rejection, work, hit, get, complete, check email, update, answer, response, attention, reason, one, message, would, instructions, respond, contact, review, say, avoid, receive, survey, worker, submit, still, days, like

more. The topics (i.e., problems) associated with the task design phase discovered in the reviews are explained in the following:

(1) Malfunction: Implementation Errors Workers complain about technical problems with the MTurk website or interfaces requiring them to use and work on tasks, which can be frustrating and make it hard for them to complete tasks efficiently. Some issues lie in the slow reaction times of the interfaces, which can lead to

breaking the loading request, difficulty accessing tasks due to error messages, a broken link, or bugs in websites or interfaces that complicate successful work and payment. Figure 3.3 shows two review examples discussing the malfunction interfaces problem.

[Malfunction: Implementation errors]

- **Review 1:** "Survey link goes to 404 page (hit attempted 3/30/17),

  I suspect there is something wrong with the URL but I couldn't find a way to edit it to get to work. Reported to Amazon and returned."

- **Review 2:** "broken, dead link (4/28/18),

  The link opens to a page with a Start button. When I clicked that button I got a. popup with this:,

  'This experiment is not currently available.',

  I've seen this exact page before, but I haven't posted a review about it here so it must have been for a different requester - maybe someone who is using the same broken template. In any case, the survey is not accessible so I reported it to Amazon."

FIGURE 3.3: Two examples of the reviews that are classified as they mainly contain stories of "Malfunction: Implementation errors" problem.

(2) Malfunction: Failed Completion For some tasks on MTurk, a completion code is a unique code generated when a worker completes a HIT. The platform uses this code to verify that the worker has completed the task and ensures that the worker gets compensated for their work. However, workers were reporting technical issues with the MTurk platform that resulted in failed completion codes or submission procedure due to timeout problems, which can lead to payment issues and be frustrating for the worker. If the submissions do not meet the required standards, the requester may manually return or reject the worker's completion code. In such cases, workers may contact MTurk customer service or requesters for assistance. Figure 3.4 shows two review examples discussing the malfunction completion process problem.

(3) Workload: Bad Time Estimation In crowdsourcing, estimating the required time to complete a task can be essential for both workers and requesters, and failure in accurate estimation can lead to problems for both parties. An underestimate of the time required can also be problematic, as it may cause the worker to feel rushed or stressed to complete the task within the given time frame. Besides, an overestimate of the time required for a task can lead to problems with project planning and budgeting. If the task takes longer than expected, it can lead to delays and additional costs. In practice, workers complain that requesters often fail to

[Malfunction: Failed completion]

- **Review 1:** "No survey code given at the end of survey.

  'We thank you for your time spent taking this survey. Your response has been recorded.'

  Unable to submit the HIT, notrecommended"

- **Review 2:** "broken hit, no response.

  HIT itself is fine, but there is no completion code box to put the code once you have finished. The HIT cannot be submitted without the code, so you will be forced to return. I will update when I get a response

  EDIT: No response was received."

FIGURE 3.4: Two examples of the reviews that are classified as they mainly contain stories of "Malfunction: Failed completion" problem.

accurately estimate the required time for the tasks due to a lack of understanding or experience on the requester's part. This may lead workers to fail complete the task in the given time and nor get rewarded. Figure 3.5 shows two review examples discussing the bad time estimation problem.

[Workload: Bad time estimation]

- **Review 1:** "way too long for the pay Hit done 4/26/17,

  Well, the time estimate in the title was correct. Unfortunately This included a 4-minute video, a page of reading, and a whole lot of bubble questions.

  Too much for the pay. Stay away."

- **Review 2:** "Way too long for the pay.

  Lots of video and audio to watch and click. Not hard, but just never seems to end. Bad pay."

FIGURE 3.5: Two examples of the reviews that are classified as they mainly contain stories of "Workload: Bad time estimation" problem.

(4) Workload: Too High Work Effort Similarly, workers complain about underestimating the required effort, especially for writing tasks, according to the reviews. Since the workers are often distributed and may have varying skills, experience, and motivation levels, estimating crowdsourcing effort can be particularly hard for requesters. Besides, it is challenging to estimate the required effort due to the complexity and difficulty of tasks that can range widely, even within the same crowdsourcing platform. Generally, the issue of inaccurate effort estimation in crowdsourcing can result in various negative outcomes, including missed

deadlines, unsuccessful high-quality submissions. Figure 3.6 shows two review examples discussing the too high work load problem.

[Workload: Too high work effort]

- **Review 1:** "requires you to install software that tracks your internet activity and shopping, Install internet browsing tracker for

  $5. The HIT is to install an internet browsing tracker plus a long survey.

  They will track your internet browsing and shopping habits. Who does this for money? Who would let this happen?"
- **Review 2:** "writing, not going to do it threw it back -- has an open-ended writing section (at least one) of the sort I refuse to do.

  At least, not for pennies."

FIGURE 3.6: Two examples of the reviews that are classified as they mainly contain stories of "Workload: Too high work effort" problem.

(5) Violation of Terms of Service Moreover, workers report that requesters attempt to ask for workers' personal information, such as real names, addresses, and social media IDs, or require them to download an app and sign up with their private email addresses. However, according to MTurk's terms of service, requesters are not allowed to ask workers for personal information or email addresses as part of a task. This type of request may be considered a violation of the MTurk terms of service (ToS), which directly declare that requesters should not request personal information from workers or engage in activities that may violate their privacy. Figure 3.7 shows two review examples discussing violation of TOS problem.

[Violation of terms of service]

- **Review 1:** "This requester has a history of TOS violations (asking for personal info like an email address).

  If you decide to try this, don't give any personal info and if it asks for any, report the hit and return it."
- **Review 2:** "Violates TOS

  They want turkers to register at a different website. They are collecting e-mail used to sign up. Collecting personally identifiable information. Violates TOS."

FIGURE 3.7: Two examples of the reviews that are classified as they mainly contain stories of "Violation of terms of service" problem.

(6) Underpaid Tasks Workers on MTurk also face the problem of underpaid tasks, which pay less than what workers believe is fair or appropriate for the effort and

time required to complete the task. Underpaid tasks can be frustrating for workers and may lead to low morale and a poor experience on the platform. A variety of factors can contribute to underpaid tasks on MTurk. The task requesters determine the pay rates for tasks on the platform, and some requesters may not offer fair or competitive pay rates. Furthermore, the complexity and difficulty of tasks vary greatly, and workers may believe they are not fairly compensated for more complex or time-consuming tasks. Figure 3.8 shows two review examples discussing underpaid tasks problem.

[Underpaid tasks]

- **Review 1:** "very bad pay for a writing hit

  UNDERPAID WRITING, Writing tasks this involving must pay a lot more

  Avoid this cheap requester who does not fairly price writing HITs."

- **Review 2:** "horribly underpaid -May 14th 2017

  Survey about opinions on government programs, corporate banks and regulations. Some light writing involved.

  Way underpaid. 2 Day AA."

FIGURE 3.8: Two examples of the reviews that are classified as they mainly contain stories of "Underpaid tasks" problem.

Task Evaluation Problems

The topic model also identified a topic in the reviews where workers discussed the challenges they faced after submitting their work, specifically due to the inadequate assessment of their submissions by the requesters. Complaints about the task evaluation phase mainly revolved around unfair rejections of submissions, which potentially damages the workers' reputations on the given platform. The topic pertaining to the task evaluation phase of crowdsourcing, extracted from our data analysis, is described in detail below:

(7) Unfair Rejection Workers complain in the reviews that task requesters are privileged to reject a worker's work without a clear justification or explanation and refuse to pay compensation to the workers. Although, they can keep the rejected results and profit from them. Requesters often avoid paying attention to workers' inquiries about rejection and responding to their messages. It can cause workers frustrations and a sense of unfairness, as they believe they are putting effort into completing tasks but are not fairly compensated for their efforts. Figure 3.9 shows two review examples discussing unjustified rejections problem.

[Unfair rejections]

- **Review 1:** "Will reject with no possibility of resolve and not respond back

  I was rejected for being a male and doing the survey. I did not read anywhere that males are not allowed and he rejected my work.

  STAY AWAY from this requester or you will get rejected as well and will not respond to your messages."

- **Review 2:** "Will not respond back to emails

  Rejected hit saying I failed to provide valid responses. I tried to contact her to get a better explanation and she has failed to respond back."

FIGURE 3.9: Two examples of the reviews that are classified as they mainly contain stories of "Unfair rejection" problem.

The following section aggregates the findings of both literature review and the data analysis approaches and provides a comprehensive in-depth discussion of the results.

3.4 Evaluation

Given the outcomes of the literature review and the empirical analysis on the data crawled from the Turkopticon platform, we first aggregate the findings of both approaches to form a set of problems discussed the most in the literature and among workers on forums (RQ1). We then assess the dominance of each problem compared to the entire findings to obtain insights into their importance (RQ2).

Figure 3.10 summarizes the aggregation of problems found from both literature review and the data analysis together with the distribution of challenges found in literature and of probabilities in the topic model. This section describes the result of the aggregation and addresses the research questions in details.

Problems

In light of RQ1 (Section 3.1), we compared the data analysis outcomes to those of the literature evaluation to collect the discussed problems among workers in practice and researchers from a more theoretical point of view (Fig. 3.10). Here, we discuss the first research question as well as the pertained findings resulting from our work.

RQ1. Workers' Challenges in Crowdsourcing Phases We discovered a total of 14 problems (Fig. 3.10) that are created within all stages of the crowdsourcing process: five challenges pertain to the task design, three to the task operation, and

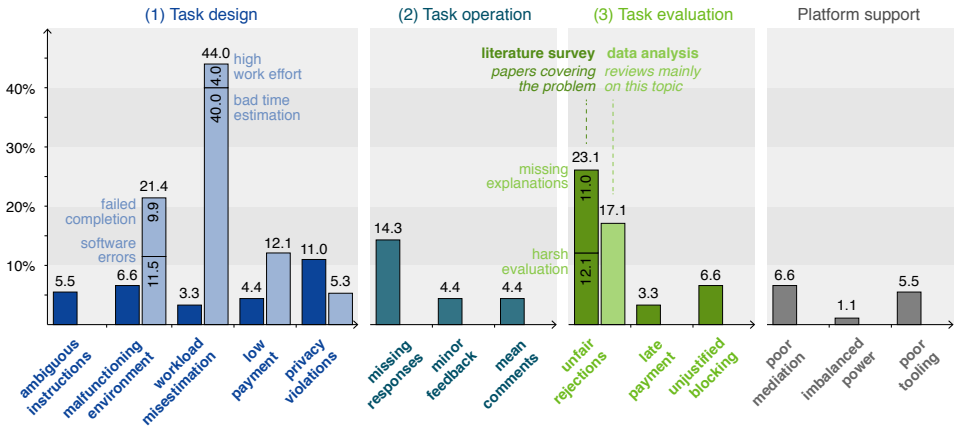


FIGURE 3.10: Distribution of the 14 found problems that workers face in the three phrases of crowdsourcing processes and with the associated platform support. Dark bars: Relative frequency of each problem in the reviewed literature. Light bars: Topic modeling probability of each problem in the data analysis (Nouri et al., 2020).

three to the task evaluation. Three problems are also rooted in the way the platform intermediates the process (details on the problems in Sections 3.2 and 3.3).

Here, we summarize five issues related to task design identified through both literature review and data analysis: (1) *ambiguous instructions*, where workers struggle to comprehend the task and the expected solution; (2) *malfunctioning environment*, where technical issues in the requester-designed task interfaces hinder successful work; (3) *workload misestimation*, where an imbalance between work effort and estimated time leads to unsuccessful attempts at task completion. This issue might be reflected in (4) *low payment* (i.e., a low hourly payment ratio), where workers are not fairly compensated for their efforts to complete the task. Finally, we also discovered the problem of (5) *privacy violations*, where requesters breach the platform's terms of service by soliciting workers' personal information in various ways.

The three problems pertaining to the task operation can be outlined as (1) *missing responses* where requesters pay minimal attention to respond to workers' inquiries related to diverse matters such as tasks, solutions, and rejections. They also often provide only (2) *minor feedback* to workers' results, decreasing the opportunity to improve their quality and resubmission. Moreover, some requesters behave unfriendly and give (3) *mean comments* on the worker's results.

Moreover, the three problems related to the task evaluation can be summarized as (1) *unfair rejections* where, on the one hand, requesters or the systems set for result evaluations reject workers' results via harsh evaluations. On the other hand, they often provide no or unclear rejection reasons to the workers, which disappoints them. Another problem is (2) *late payment*, where workers must wait a long without being aware of the time requesters will take to pay the accepted results. The problem of (3) *unjustified blocking* happens when there are disagreements be-

tween requesters and workers about the provided arguments by requesters. In this case, workers insist on discussing the subject with the requesters and get blocked by requesters.

Finally, the three problems that lie in the way the platform intermediates the process are briefly explained as (1) *poor mediation* where platforms intermediate the conflicts between requesters and workers poorly, especially in case of rejections, payments, and other unfair situations based on workers' judgments. In general, there is (2) *imbalanced power* between requesters and workers in that requesters are unrestricted to reject results and not pay the workers even unfairly. Moreover, platforms partially have (3) *poor tooling*. For instance, there is a lack of automated mechanisms and proper quality control on workers' or requesters' reputations, task payments, and term-of-service violations. Besides, requesters' contact information may not be available to workers who do their tasks.

Although the data analysis provided us with a clear distinction of the dominance of the challenges that workers face in the marketplace, we discovered that all problems found in the empirical analysis are already discussed in the literature, which indicates that they have been known to the community, yet not effectively resolved.

Problems Importance

In light of RQ2, we counted the number of times each problem was mentioned or discussed in the literature and then compared the relative percentages of these problems with the problem probabilities achieved from topic modeling. Here, we review the second research question and describe the findings resulting from our work in detail.

RQ2. Importance of the Challenges in the Literature and Data The literature discusses the different phases somewhat likewise, with major problems being *missing responses* (14.3%) during task operation and *unfair rejections* in the task evaluation (12.1% *harsh evaluation*, 11.0% *missing explanations*). Such rejections are also covered significantly in Turkopticon reviews (17.1%). However, the major problems discussed in reviews pertain to the task design where workers complain about *workload misestimation* (44.0% in total). They also report *low payment* (12.1%), along with *malfunctioning environments* (11.5% *failed completion*, 9.9% *software errors*) in the reviews. Workers also shared narratives of *privacy violations* (5.3%) regarding the platform's terms of service.

Problems involving task design and task evaluation are well observable in both sources. On the contrary, we could not explicitly identify problems concerning platform support and task operation in the data. One factor is topic modeling's poor capability to differentiate between discrete and fine-grained topics. Conse-

quently, our data analysis may have potentially overlooked certain issues. This oversight occurred not because these problems were entirely absent from the reviews, but rather because they were only addressed in a few of them. For instance, some people bring up issues with platform mediation as well as non-responsive requesters. Talking exclusively about the latter, though, seems uncommon.

Our data analysis shows that the task design is the primary source of issues from the workers' perspective. This is consistent with our reasonable assumption that the data reflects what most irritates the workers and motivates them to report to others.

3.5 Conclusions

Our primary focus is on improving the crowdsourcing process through the application of computational methods. This chapter presents a combination of literature review and data analysis to identify the challenges faced by workers when working with requesters in various aspects of crowdsourcing tasks such as design, operation, and evaluation. The literature review encompasses both theoretical findings and insights gathered from interviews and questionnaires that specifically address the issues faced by workers. In the data analysis, we employ topic modeling techniques to extract relevant problems from workers' complaints found in an online discussion forum. Additionally, we provide the underlying corpus as a valuable resource for further research within the community. By merging insights from both the literature and data analysis, our aim is to bridge the gap and facilitate the exchange of knowledge between the research fields of human computation and natural language processing.

During our research, we had the opportunity to engage in discussions with a manager from MTurk, who informed us about the platform's efforts to address the challenges encountered by both requesters and workers. As a result of these initiatives, MTurk has introduced several new policies. For instance, the platform now displays a requester's "acceptance rate," which indicates the proportion of tasks the requester has accepted in the past. This feature aims to assist workers in identifying reliable requesters. Furthermore, MTurk provides a wide array of general guidelines and policies for both requesters and workers, including task design templates and terms of service. Despite these measures, our study reveals that numerous problems persist within the crowdsourcing ecosystem.

Upon analyzing the identified problems, it becomes evident that the majority of them stem from deficiencies in task design and communication throughout the task's execution and evaluation. Task design holds significant importance in crowdsourcing processes as it directly impacts the quality of the obtained results (Kittur, 2010). Issues arise when requesters fail to appropriately break down complex tasks into manageable components and provide fair compensation for the effort invested.

Insufficient communication negatively impacts worker satisfaction, reputation, and accomplishments (Bederson and Quinn, 2011, Boons et al., 2015).

Additionally, errors in the implementation of the task environment lead to wasted time for workers due to unsuccessful submissions. Some requesters are also reported to violate workers' privacy by disregarding the terms-of-service of the crowdsourcing platform being used. Unfair rejections without explanation emerge as the predominant issue in task evaluation, and the literature extensively discusses communication problems between requesters and workers, exacerbated by inadequate platform support. Considering that workers are the primary contributors to the outcomes, enhancing the design and communication processes has the potential to enhance the overall quality of crowdsourcing.

We reexamined our findings in consultation with social scientists involved in our research project. They proposed that the inclusion of personal information requests in task design may be influenced by traditional selection mechanisms, where employers often adhere to stereotypical notions of an ideal employee based on specific requirements (Acker, 1990, Van der Lippe et al., 2019). Consequently, requesters might attempt to bypass the terms of service that prohibit them from soliciting personal information from workers. The example reviews presented in Section3.3 provide some support for this hypothesis; however, further investigation is necessary to substantiate it in future studies.

Given the rising popularity of crowdsourcing, it is imperative to make improvements to ensure a fair and effective process for both requesters and workers. While examining the problems from the requesters' perspective would provide valuable insights, the current literature and available data for this purpose are lacking.

In this thesis, our aim is to enhance the crowdsourcing processes by providing technological support in task design. We envision the development of an automated assistant system that assists in improving the clarity of task descriptions. We believe that this is a crucial step in helping requesters, especially those who are new to crowdsourcing, learn from the mistakes of previous requesters. By creating clearer task descriptions, requesters can expect to receive higher-quality results from workers. Based on our findings, we firmly believe that this approach has the potential to address numerous challenges prevalent in current crowdsourcing processes.

Chapter 4

Task Clarity Assessment

Crowdsourcing is growing extensively and and has proven to be highly advantageous for both organizations and individuals (Howe et al., 2006). Crowdsourcing marketplaces enable on-demand access to mixed human input, leading to cost-effective solutions and services. This flourishing paradigm delivers the potential to harness the knowledge, capabilities, and creativity of a crowd for problems that demand human intelligence. As explained in Chapter 2, the general crowdsourcing process has three main phases (Nouri et al., 2020): (1) *Task design*, where requesters write and post their task descriptions on a crowdsourcing platform. (2) *Task operation*, where workers decide to accept tasks and then submit their solutions. Workers may inquire about task details, and requesters may answer their questions or give feedback on the solutions. And (3) *task evaluation*, where requesters choose to accept solutions and to pay workers.

The main focus of our thesis is derived from the findings of our previous study, as discussed in Section 3.4. This study revealed that a considerable portion of crowdsourcing challenges, particularly from the workers' perspective, originates from the task design phase. Requesters often struggle to accurately estimate the effort and time required to complete a task, resulting in issues related to fair compensation, the provision of a conducive working environment, and ambiguous task descriptions. Additionally, the predominant focus of research in the field of crowdsourcing has centered around the quality of solutions provided by crowd workers (Kittur et al., 2013). The recognition of low-quality solutions as a significant impediment to harnessing the full potential of crowdsourcing has been emphasized (Weld et al., 2015), and unclear task design has been identified as a critical factor negatively affecting the quality of crowd work (Kulkarni et al., 2012, Gadiraju et al., 2017, Wu and Quinn, 2017, Manam and Quinn, 2018).

Therefore, among all the challenges identified in the previous study, our primary objective in this thesis is to improve task description quality. We acknowledge that ensuring clear task descriptions is essential for achieving high-quality task design.

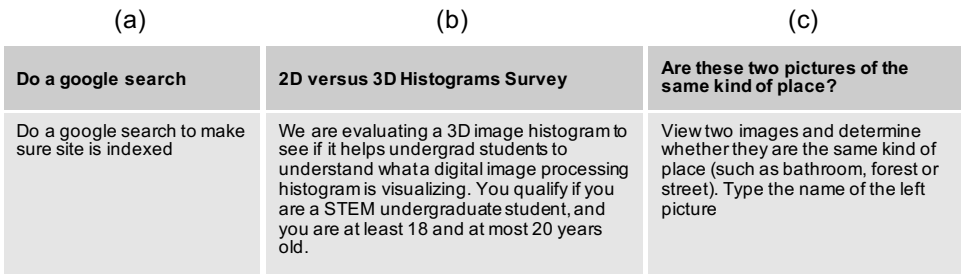


FIGURE 4.1: Example crowdsourcing task descriptions from the dataset introduced in Section 4.3.

As shown in Figure 4.1, a typical task description consists of a title and a set of instructions that need to be easily understandable. It is important for task descriptions to clearly define the expected solution and provide guidance on how it should be achieved (Kittur et al., 2008, Grady and Lease, 2010, Alonso and Baeza-Yates, 2011, Franklin et al., 2011, Finnerty et al., 2013, Manam et al., 2019). These descriptions directly impact how workers comprehend and choose tasks (Little et al., 2010, Schulze et al., 2011), influencing their participation (Khanna et al., 2010) and task completion rate (Chen et al., 2011), as well as their acceptance rate, reputation, and payment (Silberman et al., 2010b), despite the time and effort invested in completing tasks (Manam and Quinn, 2018). The quality of task descriptions has a significant impact on the quality of results obtained and the satisfaction and confidence of workers (Finnerty et al., 2013, Wu and Quinn, 2017). In fact, Khanna et al. (2010) provided evidence that task clarity, as achieved through clear descriptions, improves the usability of crowdsourcing, particularly when engaging low-income workers.

Hence, writing clear task descriptions is crucial in the context of crowdsourcing. However, previous research on crowdsourcing has consistently highlighted the challenge of vague task descriptions (Fowler Jr, 1992, Khanna et al., 2010, Chandler et al., 2013, Gadiraju et al., 2017, Gaikwad et al., 2017, Wu and Quinn, 2017, Nouri et al., 2020). On one hand, requesters are expected to provide comprehensive information necessary for task completion, including the required resources, step-by-step instructions, and desired solution format. This can be particularly challenging for requesters with limited crowdsourcing experience, especially when dealing with micro-tasks that target a diverse pool of workers from different cultural backgrounds, skill sets, and educational levels (Ipeirotis, 2010). On the other hand, crafting clear and understandable task instructions is a complex task due to the subjective nature of requester's wording and the inherent ambiguity of natural language (Franklin et al., 2011). Consequently, workers may interpret the instructions and requirements differently, leading to potential misunderstandings.

When examining a simple task description example shown in Figure 4.1(a), it becomes apparent that there are clarity issues present. While it can be assumed that

the "site" is mentioned in the description, the specific requirements for ensuring that the site is "indexed" may be unclear to workers. Furthermore, the technical term "indexed" itself may lack precise definition. Similarly, the longer task description depicted in Figure 4.1(b) also exhibits clarity flaws, including a lack of clarity regarding the overall expected solution (i.e., what needs to be provided for approval and the method of doing so). On the other hand, the task description illustrated in Figure 4.1(c), while not perfect, provides clear instructions on what needs to be done and how, utilizing simple and easily understandable language without any complex wording.

We claim that addressing the dual problem of accurately and comprehensively describing necessary information in task descriptions can derive significant advantages from technological support for requesters. This support should enable them to enhance the clarity of their task descriptions while ensuring that all essential details are included. Interactive tools that provide automated assistance in improving the quality of descriptions would be highly valuable to task requesters. However, the development of such tools has been hindered by the absence of effective computational approaches for assessing the clarity of descriptions.

In Section 2.3, we conducted a review of studies that employ natural language processing techniques to examine text clarity. Additionally, we delved into the task clarity, where we explored the creation of workflows and models designed to improve the clarity of tasks in crowdsourcing marketplaces. In contrast, the computational approach presented in this chapter facilitates an automated tool that brings about a quicker and more cost-effective workflow for the requester by completely removing the workers from the process. The intended workflow remains unaffected by various challenges discussed in prior research (Nouri et al., 2020). These challenges encompass issues like subpar results from workers, difficulties stemming from strained requester-worker relationships, and the absence of an adequate feedback system in the process. However, if required, the approach can still be complemented by subsequent processes that take into account the workers' opinions.

This chapter is dedicated to addressing the central focus of this thesis—the problem of unclear task descriptions in crowdsourcing, and the entire work presented here is published in (Nouri et al., 2021a). Section 4.1 provides an overview of the methodology that has been adopted to explore the feasibility of using computational methods to tackle this issue. Subsequent sections delve into each step of the approach, offering a comprehensive understanding of the process. Section 4.2 elaborates on the computational approach adopted for this assessment study. The process of creating the required dataset for the approach is outlined in Section 4.3. Following that, Section 4.4 provides intricate details about the experiment conducted to investigate whether natural language processing techniques are beneficial for assessing task clarity flaws. Finally, the results of the experiments are

discussed, providing answers to the research questions. In the following, we introduce the approach of the computational assessment study on crowdsourcing task description clarity.

4.1 Approach

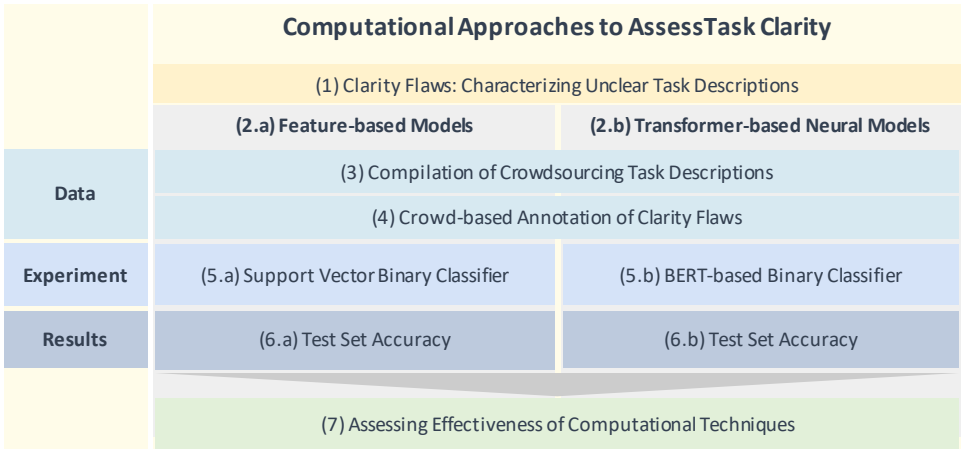


FIGURE 4.2: Overview of the approach adopted to evaluate the computational techniques for clarity flaw assessment.

In this chapter, our objective is to investigate the extent to which natural language processing techniques can be utilized to evaluate the clarity of crowdsourcing task instructions. To achieve this, we address the following research questions:

> **(RQ1)** How effectively can task descriptions' most common clarity flaws be identified automatically?
>
> **(RQ2)** What textual properties render a task description unclear with respect to the defined flaws?

Figure 4.2 illustrates our contribution to the state of the art in supporting task design, with more detailed information available in Section 2.3. Our objective is to assess the feasibility of computationally evaluating clarity flaws in crowdsourcing task descriptions. To achieve this, we follow an approach outlined as follows: We define a set of common clarity flaws in task descriptions (Fig. 4.2(1)), hypothesizing that these flaws contribute to unclear task instructions. We then develop two natural language processing approaches for computational task clarity assessment (RQ1): a) Linear feature-based models (Fig. 4.2(2.a)) to gain insights into the impact of textual properties (RQ2). b) Transformer-based neural models (Fig. 4.2(2.b)) for task clarity assessment. Afterwards, we utilize a dataset

consisting of actual micro-task descriptions (Fig. 4.2(3)) and extend it with crowd-based annotation to label clarity flaws (Fig. 4.2(4)). Utilizing the annotated dataset, we assess each method through experiments in two binary classification setups, namely feature-based support vector machines (Fig. 4.2(5.a)), and BERT-based models (Fig. 4.2(5.b)). We then compare the effectiveness of the approaches by calculating the test set accuracy (Fig. 4.2(6.a) and (6.b)). Eventually, we draw conclusions whether natural language processing models can successfully assess clarity flaws in task descriptions (Fig. 4.2(7)).

Through the adoption of this approach, our objective is to make a meaningful contribution to computationally evaluating task description clarity flaws. Altogether, the major contributions of this chapter are:[1]

- A dataset: we create a dataset with 1332 real-world task descriptions annotated by crowd workers for study task clarity in crowdsourcing.
- Computational methods: we develop a feature-based and a neural approach for the computational assessment of task clarity.
- Extensive empirical insights into task clarity assessment: we collect results to answer what helps in the computational assessment of task clarity and to what extent.

4.2 Computational Methods for Assessing Task Clarity

The main objectives of this study are to examine, based on the plain text task descriptions, how well computational techniques can assess the defined clarity flaws and what textual attributes of task descriptions indicate their clarity flaws. Therefore, we begin by providing a brief overview of the concept of task description clarity in the context of crowdsourcing. We explore the challenge of designing textual descriptions that encompass all the necessary information for successful completion of a crowdsourcing task, while also being easily understandable for a diverse pool of crowd workers. Following that, we conduct a synthesis of relevant literature to identify common flaws associated with task clarity and provide a characterization of unclear task descriptions. We then outline the two techniques we investigated for our goals: traditional feature-based and state-of-the-art neural models. Both are motivated and thoroughly introduced in this section.

Clarity Flaws: Characterizing Unclear Task Descriptions

Task clarity refers to a twofold attribute of crowdsourcing task descriptions, which not only impacts the comprehensibility and completeness level of the instructions written in natural language but also defines the extent to which requesters specify

[1]The Data and experiment code are available here: `https://osf.io/m8njv/`

the required details in instructions for receiving a high-quality solution to the task. Task clarity in crowdsourcing marketplaces is influenced not only by the establishment of participation criteria and eligibility constraints for tasks, such as reputation, experience, demographic variables, and language proficiency, but also by the task design employed by requesters. The way tasks are structured and presented plays a crucial role in ensuring clear instructions and expectations for workers.

In particular, issues with task clarity may appear due to the inexperienced requesters, who may lack an awareness of the variety among target workers in terms of their educational background, skills, demographics, and culture. Likewise, they may lack adequate understanding of the significance of an effective task design and its direct impact on the quality of worker submissions. Unclear task descriptions can cause inaccurate answers from workers, leading to assignment rejection and distrust between requesters and workers.

In order to investigate the various dimensions of task description clarity flaws that impact workers' understanding of tasks, we conducted research on the literature that specifically discussed crowdsourcing task description clarity. Different researchers have discussed challenges about task clarity and studied dimensions that lead to descriptions being perceived as incomplete or unclear. Among these, Gadiraju et al. (2017) discussed *goal clarity* and *role clarity* as the main aspects of micro-task descriptions. These terms refer to the expected solution, and the way crowd workers should do the work, respectively.

Besides, Wu and Quinn (2017) presented the concept of *descriptive* and *prospective metrics* of task descriptions. Descriptive metrics are of particular interest in determining the clarity flaws of task descriptions. They include 1) the vocabulary or language used to describe the task, 2) the specification of the data that is expected to be delivered by crowd workers, 3) the order of the steps that should be followed in a task, and the solution to be submitted. Prospective metrics refer to the more subjective task properties relating to the workers' personal feelings. Such metrics play a role in workers' trust, conviction, and forecast of results rather than their general understanding of the task, which is affected directly by task descriptions. [2]

Eventually, other information attached to best practices for a clear task instruction contains the interface on which the work should be completed, the expected solution format, and the specification of acceptance criteria (Wu and Quinn, 2017).

Based on these findings, we curated a comprehensive set of clarity flaws (published in (Nouri et al., 2021a)) that cover both readability and understandability, which are generally significant properties of text clarity. Additionally, we included the characteristics of clarity concerning crowdsourcing task descriptions that have been discussed in the literature. However, we omitted the prospective metric mentioned by Wu and Quinn (2017) from our set due to its subjective nature, as it relies

[2]Rare exceptions may appear for tasks where the specific object of investigation is especially known to a worker. Nevertheless, the general form of a description is of a particular interest rather than its specific object of investigation.

on workers' personal feelings rather than their comprehension of a task according to its description.

Here, we outline the set of clarity flaws that describe the ground for annotation guidelines in creating the required dataset in Section 4.3. We here suggest modeling clarity by estimating the following clarity flaws. The *description unclear* can be considered as an overall unclarity of descriptions, while the remaining unclarity flaws refer to the dimensions of ambiguity in task descriptions:

1. **Description unclear.** The task is vague, i.e., it is not entirely comprehensible what the task requester expects as the desired solution and/or how one can create this desired solution successfully. This clarity flaw refers to overall unclarity.

2. **Difficult wording.** The wording, phrasing, and grammatical constructions used to write the task descriptions are not fully understandable (Wu and Quinn, 2017).

3. **Important terms undefined.** The potentially important terms to adequately understand the tasks are not defined sufficiently. This clarity flaw dimension refers to the terminology used for the description (Grady and Lease, 2010).

4. **Desired solution unspecified.** The requester did not sufficiently explain in detail the character of the desired solution expected in response to a task. This clarity flaw refers to the goal clarity aspect (Gadiraju et al., 2017).

5. **Solution format unspecified.** The task requester did not sufficiently specify the expected format of the desired solution, e.g., a piece of text, answering multiple-choice questions, reactions in social media, sharing some content, etc. This clarity flaw refers to the necessary detailed information related to goal clarity (Wu and Quinn, 2017).

6. **Steps unspecified.** The requester did not sufficiently specify which steps workers must follow to complete a task and achieve the expected solution in the desired format. This clarity flaw becomes significant when workers must go through multiple specific steps on a secondary platform to complete the task and submit the solution. This clarity flaw refers to the role clarity aspect (Wu and Quinn, 2017).

7. **Resources unspecified.** The requester did not sufficiently specify the required resources, such as data, tools, links, websites, etc. Similar to the previous clarity flaw, this clarity flaw becomes noticeable when workers must complete the task on another platform using resources that are not performable on the crowdsourcing platform.

8. **Acceptance criteria unspecified.** The requester did not sufficiently specify the criteria for evaluating and decision-making whether the submitted solutions should be accepted and rewarded. This clarity flaw refers to the helpful details for workers to estimate how much time is required to complete the task and how much the effort is remunerated (Kulkarni et al., 2012).

In the following, we introduce two techniques that we employed to computationally assess the clarity of task descriptions.

Neural and Feature-based Clarity Assessment: Estimating Effectiveness

In this study, we employ two techniques to compare their effectiveness in assessing clarity. Figure 4.3 illustrates the process through which we apply these techniques to address the research questions. To investigate Research Question RQ1 from Section 4.1, we utilize the labeled dataset which will be introduced in Section 4.3. Firstly, we employ linear SVM classifiers based on six feature types, as they have demonstrated effectiveness when data is restricted. Secondly, we rely on transformer-based neural models, which have shown superiority in various natural language processing studies. The outcomes of these techniques are then compared to estimate the effectiveness of the computational approach in assessing description clarity.

In the following, we provide an overview of the feature-based technique, introducing the six feature types used to build classifiers. The results of this technique address RQ2. Subsequently, we explain the neural approach used for task description clarity assessment.

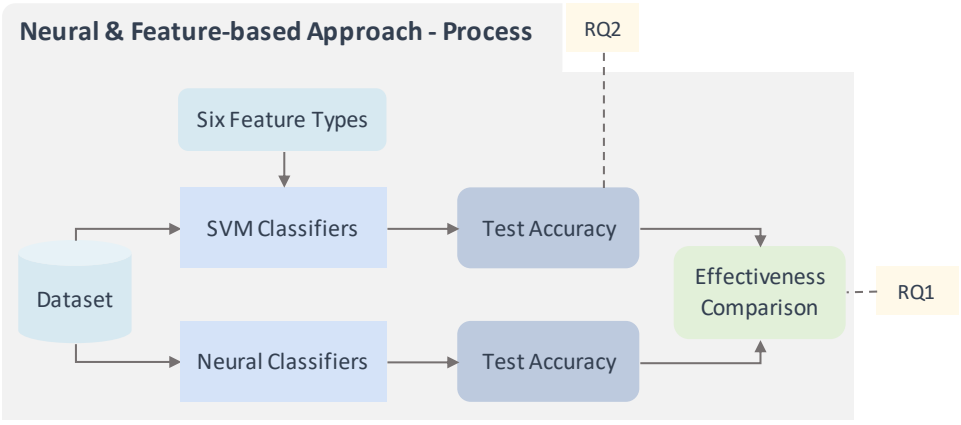


FIGURE 4.3: The procedure for evaluating the effectiveness of feature-based and neural methods in assessing the clarity of task descriptions.

Feature-based Approach Feature-based classifications can provide an in-depth understanding of the textual properties that help assess each clarity flaw, enabling us to address RQ2. Our linear SVM classifiers are particularly built upon the following six feature types that are published in (Nouri et al., 2021a). These feature types include both standard features that have usually been utilized in natural language processing as well as the clarity flaw-specific feature type that we created based on the well-known elements of task descriptions (Gadiraju et al., 2017):

1. **Content.** Content is essential in numerous text classification tasks. Consistent with standard practice, we study the impact of content-related properties through term frequency-inverse document frequency (TF-IDF), considering all lower-cased tokens as 1- to 3-grams, including stop words.

2. **Length.** To analyze the correlation of clarity with length property, we incorporate 26 features that show the extent of a task description. They reflect the numbers of all unique words, words, characters, non-whitespaces, whitespaces, punctuation marks, digits, letters, fully upper-cased tokens, fully lower-cased tokens, capitalized words, phrases, and sentences. Further, we calculate the average of all counts per sentence (except for sentences).

3. **Style.** Clarity can be considered a characteristic of the style. We accordingly model style via part-of-speech as well as phrase for both 1- to 3-grams (created using the NLTK library[3]) along with characters 3-grams and the *functional words*. We consider the top-100 most frequent lower-cased words in the whole corpus for the functional words.

4. **Subjectivity.** Subjective wording of task descriptions has been highlighted to influence task clarity (Gadiraju et al., 2017). We measure subjectivity using the *Textblob* library[4], which calculates a subjectivity score, a negativity score, a positivity score, a polarity score, and an objectivity score for a given textual content.

5. **Readability.** As Section 2.3 mentions, readability measures have been utilized to assess description clarity. We consider ARI, Coleman-Liau, Flesch-Kincaid Grade Level, Gunning-Fog Index, LIX, SMOG Index, RIX, Flesch Reading-Ease, and Dale-Chall Index. All readability measures are calculated via the Pypi library[5].

6. **Flaw-specific.** In line with the clarity flaws, we assume that the task description clarity is reflected in the *completeness*, relating to resources and acceptance criteria, along with the *complexity* of phrases and terms. We introduce

[3]NLTK library's Homepage link: `https://www.nltk.org`
[4]Textblob library's Homepage link: `https://textblob.readthedocs.io/en/dev/`
[5]Pypi library's Homepage link: `https://pypi.org/project/readability/`

the following eight task-specific features to enable studying completeness and complexity. The first four are binary features measuring whether a description matches the provided regular expression.

a. *Website.* A regular expression for diverse tokens 1- or bi-grams, which refer to an online resource (e.g., "web page(s)", "webpage(s)", "site(s)", "web site(s)", etc.).

b. *Link.* A regular expression for web addresses, URLs and words such as "link".

c. *Given time.* A regular expression for token 1-grams delivers information regarding the estimated completion time, such as "2 minutes", "3 minutes 15 seconds", "5 min".

d. *Reward.* A regular expression for token 1-grams providing information regarding a task's determined reward (or bonus), such as "up to $0.45 + 50% bonus = $0.68 max", "10 cents", "avg rwrd+bns: $3.02".

e. *Entity.* All token n-grams caught by Spacy[6] as organizations, products, locations, ordinal entities, or similar.

f. *POS categories.* Frequencies of part-of-speech tags that are conceptually similar and found by the Stanford Tagger[7], such as nouns, open and close part-of-speech tags, verbs, and similar.

g. *Discrete words.* The ten discrete lower-cased 1-gram tokens (excluding stop words) are repeated most frequently and appeared either only in clear task descriptions class or only in unclear descriptions class for all dimensions.

h. *Complex words.* Two distinct scores for the complexity of token 1-grams calculated by Pypi.

Neural Approach We rely on the extensively used BERT model (Devlin et al., 2019). We explore two standard versions of pre-trained BERT, namely *Bert-base-uncased*, a case-insensitive model trained on lower-cased English text, and *Bert-base-cased*, a case-sensitive model trained on English text in its original structure. Both versions have 12 layers, 768 hidden nodes, 12 heads, and almost 110 million parameters [8].

In the following section, we provide a detailed account of the process involved in creating the dataset required to address the main objectives of this study, namely, the research questions.

[6]Spacy's Homepage link: `https://spacy.io/models/en#en_core_web_lg`

[7]Tagger's Homepage link: `https://nlp.stanford.edu/software/tagger.shtml`

[8]Pre-trained BERT models: `https://huggingface.co/transformers/pretrained_models.html`

4.3 A Dataset for Assessing Task Clarity

In order to facilitate the study of task description clarity assessment, we undertook a four-step process to create and validate a dataset. These steps were as follows: 1) The compilation of existing task descriptions, 2) The annotation of the task descriptions for the clarity flaws, 3) The consolidation of the final dataset, and 4) A fundamental correlation analysis of the clarity flaws. In the following, we clarify each step, introducing the source task descriptions, the annotation procedure, along with the resulting data distribution and correlations.

Compilation of Crowdsourcing Task Descriptions

For the data compilation, we extended the earlier published dataset of Gadiraju et al. (2017) and Difallah et al. (2015), which comprised of a total of 7007 real-world task descriptions posted on Amazon Mechanical Turk (MTurk) between October 2013 and September 2014. We know that the age of the task descriptions may affect what we observe. Nonetheless, we favored comparability to prior work over timeliness as we did not notice an essential change in the descriptions after 2014. Moreover, note that acquiring task descriptions is all but straightforward. The title, body, posted date, and other metadata are given for each task. For our study, the title, a dot (as a separator), and the body of the instructions compose the task descriptions in the dataset. The task descriptions are classified into six distinct task types, namely *Surveys (SU)*, *Content Creation (CC)*, *Content Access (CA)*, *Verification and Validation (VV)*, *Interpretation and Analysis (IA)*, as well as *Information Finding (IF)* (Gadiraju et al., 2014).

TABLE 4.1: The distribution of task descriptions across the six unique task types within the original dataset (Gadiraju et al., 2017), after eliminating almost identical entries, and within the final 50% sample that we utilized for our annotated dataset.

#	Task Type	# Original	# No Near-Copies	# Our Dataset
SU	Surveys	1200	1121	561
CA	Content Access	1008	528	264
IA	Interpretation and Analysis	1199	505	253
IF	Information Finding	1200	291	144
CC	Content Creation	1200	147	74
VV	Verification and Validation	1200	71	36
	Total	**7007**	**2663**	**1332**

While inspecting the original dataset, we noticed that the 7007 task descriptions comprised many cases that were near-copies of others regarding multiple instances of the identical task only with certain information replaced. Given that we did

not see any clarity-specific dissimilarities in these instances, we employed a semi-automatic procedure to filter out near-copies. First, we manually grouped the near-copies, and then through an automatic procedure, we randomly selected 50% of these groups, resulting in a final set of 2663 distinct task descriptions. Due to the restricted budget, we selected 50% task descriptions for manual annotation (resulting in 1332 records). To maintain the highest diversity coverage of task descriptions, we picked the sample representative concerning the task types in the filtered set. Table 4.1 displays the distribution of the task types in the resulting dataset compared to the source data. The 1332 task descriptions include 31,027 tokens (23.3 tokens per description on average) and 25,891 unique tokens.

Crowd-based Annotation for Clarity Flaws

We decided to gather annotations for the clarity flaws in the task descriptions straight from crowd workers because they eventually benefit from more clear and precise task descriptions, making their opinion decisive. Following the origin of the available task descriptions, we posted the annotation tasks on MTurk, so that the participants for our annotation tasks match the potential workers for the given descriptions in principle.

Task Design As task requesters in this setting, we made a concerted effort to avoid the clarity flaws introduced in Section 4.2 as much as possible in our task description. In our annotation task, the crowd workers were asked to evaluate a given task description from our dataset for clarity flaws in the form of a questionnaire. Figure 4.4 shows our annotation task description, created during main annotation study, that workers see in the list of available tasks on MTurk platform based on which they decide to view and eventually accept the task. Initially, the general instructions were about how to fill out the form along with a privacy guarantee, requesting the workers to imagine themselves as the worker who accepts the task with the given description. Then, each flaw was covered following the definitions from Section 4.2.

To specify a proper design for the annotation task setting, especially the adequate number of annotators, we created and deployed the annotation tasks in two main phases: first, a *pilot annotation study* where we analyzed the initial design decisions on MTurk to check whether our tasks attain the desired results with satisfactory quality; and second, the *main annotation study* where we gathered the annotations of the entire 1332 task descriptions after enhancing the task design based on the results from the pilot study.

Pilot Annotation Study We drew a comparison between two distinct annotation schemes regarding which one provides a higher inter-annotator agreement:

[TASK]

In this HIT, you see four task descriptions of past crowdsourcing tasks given to the workers. You should evaluate whether given task descriptions are sufficiently clear.

[BACKGROUND]

By working on this HIT, you support a scientific project of Paderborn and Bielefeld Universities in Germany. The purpose of our research is the improvement of crowdsourcing processes. Studies show that unclear task descriptions, created by inexpert requesters, lead to confusion about how and what to be done. It means that workers do not have enough information about the steps through which the tasks need to be completed and also about what the expected solution is. Therefore, task descriptions need to contain necessary information in order to be clear for crowd workers.

[DATA PRIVACY AND PROTECTION]

Information about data privacy and data protection: The results are anonymous, i.e. all answers are recorded and stored anonymously. Your participation is voluntary. There are no disadvantages if you do not participate. Your information will be treated in strict confidence. The anonymized data will be stored and evaluated on servers of the University of Bielefeld and the University of Paderborn, only for the purpose of the scientific study. In particular, no survey data will be passed on to commercial users or administrative authorities. The results do not allow any conclusions to be drawn about individuals.

[PAYMENT CONDITIONS]

The average time to complete this HIT is estimated 5 minutes and the reward is $0,83. Although, the time to submit the results is allotted to 20 minutes to ensure a successful submission for all participants.

[NOTICE]

As a pre-step to accept the task, you need to make sure that: First, you have read and understood the information about data privacy and data protection. Second, you accept the payment conditions.

FIGURE 4.4: The annotation task description that was finally published on the MTurk platform during the main annotation study step. This task description was shown to crowd workers on the dashboard containing the list of available tasks for workers.

(a) *binary scoring* where workers either disagreed or agreed with the given statements pointing out each clarity flaw; and (b) *5-point Likert scoring* where they declared their agreement level with the statements from "1: strongly disagree" to "5: strongly agree". We then conducted a pilot study with two bunches of 12 annotation tasks (one for binary, one for Likert). Each annotation task contained four task descriptions, conveying 48 descriptions in total. Besides the flaw assessments,

we requested the workers to provide a summary of the task description, which we relied on as a quality check to check whether the annotators paid proper attention in reading the task descriptions while declaring their opinion. We considered 8 minutes to finish each task and rewarded USD 1.32 per task to every worker.[9]

Each task was annotated by three crowd workers with 1000 and more approved tasks (HITs) on MTurk, as recommended by MTurk, to guarantee the quality of annotators' work according to their reputation.

Our analysis of the pilot study's outcome indicated that the 5-point Likert scale gave the workers more freedom to make a more accurate judgment, providing a higher agreement among the workers; for the annotated clarity flaws, we noticed the full agreement varying from 40% to 63% for the binary scheme, and from 50% to 75% for 5-point Likert scoring after removing unreliable workers' annotations. Besides, the written summaries by the annotators showed the need to filter workers more restrictively to raise reliability.

Main Annotation Study We chose to obtain exclusively so-called *master workers* having an approval rate of higher than 95% who speak English from the United States, Australia, Canada, England, New Zealand, Ireland, or India. The title *master worker* is assigned to workers based on their performance by black-box algorithms of the MTurk platform. Five annotators completed each task on average in 5 minutes with an hourly wage of about 10 USD. Figure 4.4 illustrates the task description that was ultimately created for data annotation.

We improved the quality check by substituting the summary text field with two text fields: (an optional field) for *other problems* that the workers potentially discover in the given description, and (a mandatory field) for *a brief suggestion* for enhancing the task description clarity. Our study only utilized these texts to evaluate the workers' reliability. However, analyzing workers' problems and suggestions may be interesting in future work. Eventually, we packed the 1332 task descriptions into 333 annotation tasks, each containing four task descriptions. Although there were no limitations on the number of tasks a particular worker could finish, the annotation tasks were completed after ten days, and 33 unique participants finished our annotation tasks. Figure 4.5 presents the additional instructions given to workers on top of each annotation task explaining how they should complete it.

Figure 4.6 shows an example of one out of four sections in an annotation task. Each section was called *Evaluation*, including a brief instruction and one task description, followed by eight statements with which workers expressed their agreement level for the given task description. We designed the statements (details in Fig. 4.6) covering the eight clarity flaws in order to collect annotation labels, each corresponding to one defined clarity flaw.

[9]Based on the pilot study results, which indicated that workers were able to complete the task in less than 5 minutes, we adjusted the annotation task duration to 5 minutes (as shown in Figure 4.4).

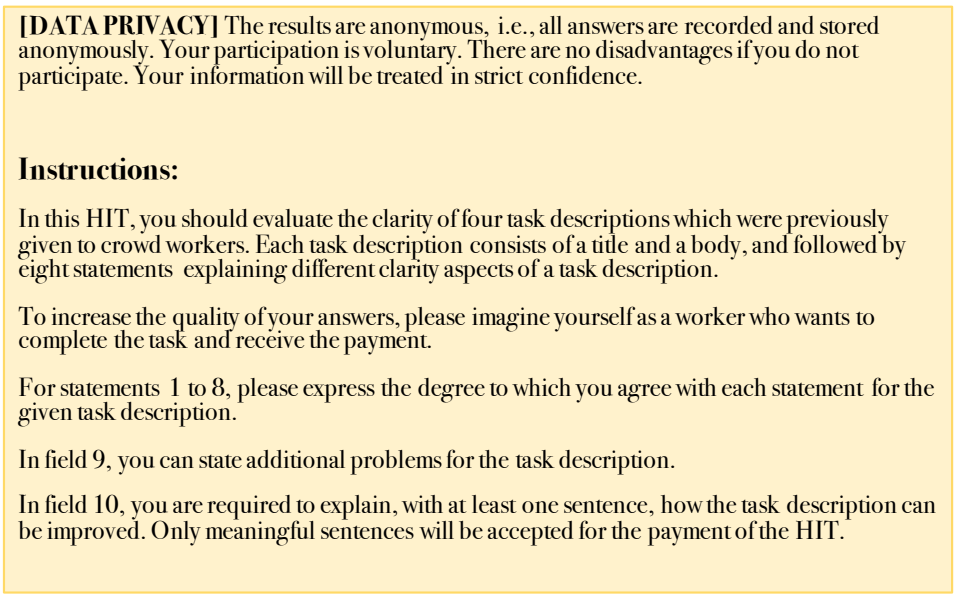


FIGURE 4.5: The additional information on our annotation task, which was launched on the MTurk platform, were displayed to the crowd workers. This information became visible when they selected the task from their dashboard list, allowing them to view the task prior to making their final decision on task acceptance. This additional instruction was given to workers at the top of the annotation task shown in Figure 4.6.

Consolidation of the Dataset

The distribution of the collected annotations for the clarity flaw "*Description unclear*" was surprisingly skewed toward "strongly disagree". We discovered three workers who had annotated more than 150 task descriptions and had chosen "strongly disagree" for the statements of more than 95% of the task descriptions. To enhance data quality, we ignored all assignments of these workers yet retained at least three annotations for all task descriptions.

We relied on the *multi-annotator competence estimation (MACE)* (Hovy et al., 2013) in order to acquire a single final annotation from the annotations left for each task. MACE was designed for crowdsourcing settings, where common inter-annotator reliability measurements such as Fleiss' κ and Krippendorff's α were not applicable due to differing annotator sets. It scores the workers' reliability based on their agreement with others and, on this basis, permits deriving one annotation for each instance.

For comparison, we also examined majority voting instead; if no majority exists, we used the rounded mean Likert scores. Still, we decided on MACE because it accounts for annotators' reliability, and the scores' distribution was also notably more balanced. The competence value of the annotators varied from 0.01 to 0.97. While the average was only 0.13, the top five had a confidence above 0.32.

Evaluation #1

Please read the following task description thoroughly and then, by carefully considering the instructions, select the statements that are true for the given task description.

Task description 1

Title: Do a google search

Body: Do a google search to make sure site is indexed

	Strongly disagree	Rather disagree	Partly disagree Partly agree	Rather agree	Strongly disagree
1. The task is clear to me, that means, I understand how to complete the task and what the desired solution is.	○	○	○	○	○
2. The wording is not easy to understand.	○	○	○	○	○
3. The potentially important terms are not sufficiently defined.	○	○	○	○	○
4. The desired solution is not explained in sufficient detail.	○	○	○	○	○
5. The format in which the solution should be submitted is not sufficiently specified.	○	○	○	○	○
6. The steps to complete the task are not sufficiently defined.	○	○	○	○	○
7. Resources that are required to complete the task are not sufficiently specified(such as data, tools, links, or websites).	○	○	○	○	○
8. The acceptance criteria for a solution to the task are not sufficiently specified.	○	○	○	○	○

9. I see other problems with the task, namely:

10. Please write in a sentence how you would improve the task description (required):

FIGURE 4.6: An example of one out of four evaluation tasks that workers on the MTurk platform were requested to complete and submit.

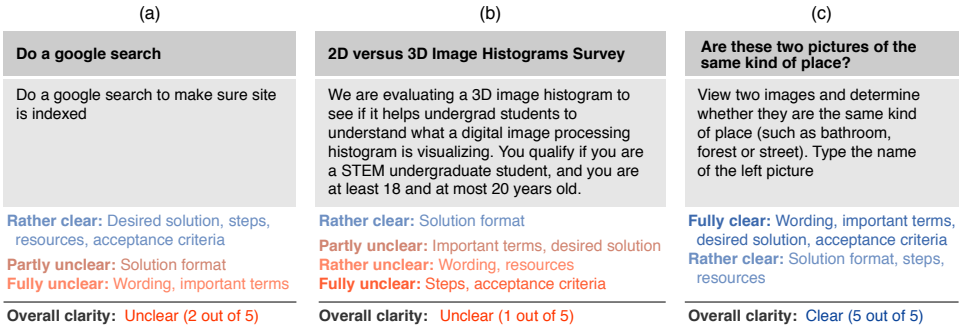


FIGURE 4.7: Sample crowdsourced task descriptions from the dataset introduced in Section 4.3. The accompanying labels below these descriptions reflect the assessments made by annotators regarding various dimensions used to gauge the overall clarity of task descriptions, averaged across these dimensions (Nouri et al., 2021a).

Figure 4.7 shows the eight final inner-annotator agreements calculated by MACE for the given task description examples (more examples in Appendix A, Section A.2), one for the overall clarity and seven for the other dimensions. According to the annotators' agreement, the wording and the important terms in the description *(a)* are completely unclear and the solution format is not sufficiently specified by its requester. In contrast, annotators' agreement indicate that only the solution format is sufficiently specified in the description *(b)* with the overall clarity lower than the *(a)*. However, the description *(c)* is clarifies what result the workers should submit and how they can achieve it.

Along with the resulting MACE Likert scores from 1 to 5, we also set a binary class for each task description, where we assume "strongly agree" (5) and "rather agree" (4) as *Positive*, and "partly agree/partly disagree" (3), and lower as *Negative*. In Figure 4.7, the task descriptions *(a)* and *(b)* are therefore classified in the negative (i.e., unclear descriptions) and the *(c)* in the positive (i.e., clear descriptions) category.

Table 4.2 illustrates the scores' distribution in the created dataset for description clarity flaws from Section 4.2. We noticed that the distribution is generally skewed slightly towards lower scores but that the whole scale is well covered in most cases. In particular, the binary classes are reasonably balanced. For the development and evaluation steps, we divided the dataset using the tasks' publication date on MTurk, to fake the idea of unseen tasks in the future for testing. The training set includes 666 task descriptions (50%) with 15,128 tokens, the validation set includes 333 descriptions (25%) with 7,821 tokens, and the test set likewise includes 333 descriptions (25%) with 8,078 tokens.

TABLE 4.2: (a) Distribution of the MACE aggregate Likert scores (Hovy et al., 2013) across all 1332 task descriptions in the dataset for each specific clarity flaw. In this context, a higher score indicates a greater extent of clarity flaw identified by the annotators. (b) The related binary scores, where ratings of 1 and 2 represent *Negative* for descriptions without flaws, while ratings of 3, 4, and 5 signify *Positive* for descriptions with flaws. It's noteworthy that the majority of values are depicted in bold.

		(a) 5-point Likert Scores					**(b) Binary classes**	
#	**Clarity Flaws**	**1**	**2**	**3**	**4**	**5**	**Negative**	**Positive**
1	Overall unclarity	**0.48**	0.18	0.09	0.20	0.05	**0.66**	0.34
2	Difficult wording	**0.42**	0.26	0.05	0.03	0.24	**0.68**	0.32
3	Important terms not def.	**0.31**	0.28	0.07	0.10	0.24	**0.59**	0.41
4	Desired solutions not spec.	0.24	**0.29**	0.16	0.22	0.09	**0.53**	0.47
5	Solution format not spec.	**0.32**	0.27	0.18	0.07	0.16	**0.59**	0.41
6	Steps not spec.	0.15	**0.35**	0.24	0.09	0.16	**0.50**	**0.50**
7	Resources not spec.	0.18	**0.32**	0.25	0.13	0.12	**0.50**	**0.50**
8	Acceptance criteria not spec.	**0.27**	0.21	**0.27**	0.13	0.12	0.48	**0.52**

Correlation Analysis

Considering the final Likert scores in Table 4.2, we conducted a correlation analysis for all eight clarity flaws to roughly assess whether they can be differentiated and to what extent they indicate that a description is generally ambiguous (overall unclarity). Figure 4.8 displays Pearson's r correlation coefficient for each flaw pair. Only several flaws correlate closely with each other; "*Difficult wording*" and "*Important terms undefined*" show the highest coefficient (0.75), which makes sense. Likewise, the flaw "*Solution format unspecified*" often highly correlates with "*Steps unspecified*" and "*Acceptance criteria unspecified*" (both 0.70). Most correlations are relatively medium, approximately between 0.5 and 0.6. A meaningful observation is that none of the seven distinct clarity flaws is closely correlated with "*Description unclear*," implying that all flaws together may add to overall unclarity. Uneasy wording shows a relatively high impact (0.59) while lacking specification of the desired solution appears slightly less critical (0.49), as long as such an inference is permitted through single correlation coefficients.

4.4 Evaluation

We now present the empirical experiments that we perform on the dataset detailed in Section 4.3 using our methodology outlined in Section 4.2. The aim is to explore the extent to which the description clarity flaws discussed in Section 4.2 can be assessed using computational techniques. In particular, we compare BERT models and feature-based SVMs in the binary classification of clarity flaws and overall unclarity. We study the extent of effectiveness that can be achieved from

	Description unclear	Difficult wording	Terms undefined	Solution unspecified	Format unspecified	Steps unspecified	Resources unspecified	Criteria unspecified
Description unclear	**1.00**	**0.59**	**0.58**	**0.49**	**0.57**	**0.53**	**0.52**	**0.54**
Difficult wording	**0.59**	1.00	0.75	0.53	0.60	0.52	0.53	0.55
Terms undefined	**0.58**	0.75	1.00	0.58	0.57	0.54	0.53	0.56
Solution unspeficied	**0.49**	0.53	0.58	1.00	0.65	0.59	0.58	0.65
Format unspecified	**0.57**	0.60	0.57	0.65	1.00	0.70	0.61	0.70
Steps unspecified	**0.53**	0.52	0.54	0.59	0.70	1.00	0.68	0.64
Resources unspecified	**0.52**	0.53	0.53	0.58	0.61	0.68	1.00	0.61
Criteria unspecified	**0.54**	0.55	0.56	0.65	0.70	0.64	0.61	1.00

FIGURE 4.8: Pearson's r correlation coefficient for each combination of clarity flaws in the dataset. The moderate correlations suggest that every unique flaw contributes, to some degree, to the overall lack of clarity (referred to as "Description unclear"). However, no single flaw is solely responsible for it (Nouri et al., 2021a).

the computational methods to assess the task clarity in light of RQ1 discussed in Section 4.1, and we examine the textual properties of instructions are helpful to determine task unclarity computationally (RQ2). We evaluate the following setups of the presented approaches along with the two baselines in our experiments:

Experiment - Feature-based Models To particularly investigate the helpful textual properties of task descriptions for computationally assessing task clarity, we employed binary SVM classifiers with the six feature types presented in Section 4.2. We utilized the *scikit-learn* library[10] to experiment with feature-based classifications. Due to the imbalanced data distribution, we randomly resampled the training set independently for each classifier. Then, we trained eight linear SVM classifiers, as mentioned, one for overall unclarity and seven for the other clarity flaws. We trained eight linear classifiers once for every six feature types separately, for each feature ablation (all feature types but one), and also for all features jointly. We optimized the cost hyperparameter per classifier on the validation set (tested range: 2^i for $-10 < i < 10$). Eventually, we calculated the accuracy score on the test set for each best validation set configuration.

Experiment - Transformer-based Models We studied the effectiveness of the task clarity flaws assessment employing eight BERT-based binary classifiers, one for overall unclarity and seven for remaining clarity flaws. In each case, we utilized the pre-trained Bert-base-uncased and Bert-base-cased models and employed the *PyTorch* library[11] to carry out the experiments on the BERT-based classifica-

[10] Scikit-learn library's link: `https://scikit-learn.org/stable/`

[11] PyTorch pre-trained BERT link: `https://pypi.org/project/pytorch-pretrained-bert/`

tions. We began with preprocessing the plain text of the descriptions operating *BertTokenizer* for both pre-trained classifiers. Then, we transformed the processed texts to the data type required for each classifier, respectively. We adjusted *BertForSequenceClassification* to the binary-label setting and employed the *BertAdam* optimizer to fine-tune the parameters with a learning rate of 2^{-5} and a warmup of 0.1. Utilizing the training set, we individually tuned the classifiers for four epochs and calculated the test set accuracy for the Bert-base-uncased and Bert-base-cased models.

Baselines We compared the introduced techniques simply to a *majority baseline*, which consistently indicates the majority training class. Thereby, we observed where learning success is obtained. The majority class is different from the training for some test sets. To determine where we can precisely distinguish classes, we also revealed the *"minority" baseline* (indicating the minority training class) below, whereas we stress that the "minority" baseline is not reasonable in practice.

Significance Tests We conducted a one-tailed independent t-test to investigate whether (a) Bert-base-uncased, (b) Bert-base-cased, and (c) the SVM with all features can estimate the task clarity flaws, significantly better than the majority baseline at $p < .05$ (marked **) and $p < .01$ (*).

Results

In this section, we compare the results of the experiments we conducted for both computational methods, a) linear feature-based support vector classifications and b) BERT-based classification, and answer the research questions of this study in detail. Finally, we conclude whether natural language processing techniques provide automated methods to assess the clarity of task descriptions relying only upon their plain text and what textual properties of descriptions are most decisive in the computational assessment.

RQ1. Effectiveness of the Task Clarity Assessment Table 4.3 presents the test set accuracy scores for both overall unclarity (referred to as *Description unclear*) and the seven other clarity flaws in task instructions. These scores are shown for both the BERT and linear SVM classifiers, considering single feature types (A_i), feature ablation ($A_{\setminus i}$), and all features (A_{1-6}). Our interpretation of these overall results helps address the first research question (RQ1).

We assessed the test set accuracy score of the BERT classifiers against the majority baseline to investigate whether the considered clarity flaws of the task descriptions can be computationally identified. We noticed that *Bert-base-uncased* and *Bert-base-cased* successfully assessed six flaws: unspecified important terms

TABLE 4.3: Test accuracy of various feature types, feature ablation, all features, and two different BERT versions in comparison to both the majority baseline and the minority "baseline" for assessing overall unclarity (referred to as "Description unclear") and the seven distinct clarity flaws. The best values in each column are highlighted in bold, and the superior feature and feature ablation are underlined. Significant improvements over the majority baseline for the *all features*, *Bert-base-cased*, and *Bert-base-uncased* settings are indicated with ** ($p < .05$) and * ($p < .01$) (Nouri et al., 2021a).

#	Approach	Description unclear	Difficult wording	Important terms undefined	Desired Solution unspecified	Solution format unspecified	Steps unspecified	Resources unspecified	Acceptance criteria unspecified
A_1	Content	0.72	0.71	0.66	0.60	0.56	0.59	**0.63**	0.58
A_2	Length	0.62	0.63	0.54	0.57	0.51	0.55	0.54	0.56
A_3	Style	0.66	0.67	0.61	0.61	0.60	0.57	0.59	**0.63**
A_4	Subjectivity	0.71	0.51	0.63	0.50	0.49	0.52	0.53	0.55
A_5	Readability	0.69	0.70	0.65	**0.63**	0.57	0.55	0.62	0.58
A_6	Flaw-specific	0.69	0.71	0.64	0.55	0.59	0.57	0.61	0.60
$A_{\setminus 1}$	w/o Content	0.72	0.72	0.67	0.62	0.60	0.58	0.60	0.61
$A_{\setminus 2}$	w/o Length	**0.74**	0.72	0.66	0.59	**0.62**	**0.62**	**0.63**	0.58
$A_{\setminus 3}$	w/o Style	0.69	0.72	0.67	0.61	0.59	0.59	0.61	0.61
$A_{\setminus 4}$	w/o Subjectivity	0.73	0.72	0.67	0.61	**0.62**	0.58	0.61	0.62
$A_{\setminus 5}$	w/o Readability	**0.74**	0.70	0.67	0.60	**0.62**	0.60	0.62	0.60
$A_{\setminus 6}$	w/o Flaw-specific	0.72	0.70	0.67	**0.63**	0.58	0.57	0.56	0.60
A_{1-6}	**All features**	0.73	0.74	*0.66	*0.61	*0.59	*0.61	*0.61	*0.62
BbC	**Bert-base-cased**	0.69	0.71	***0.69**	*0.60	*0.60	*0.56	*0.56	*0.57
BbU	**Bert-base-uncased**	0.71	0.71	*0.67	*0.62	*0.61	*0.58	*0.60	*0.55
Ma	Majority baseline	0.72	**0.75**	0.31	0.38	0.36	0.43	0.41	0.44
Mi	Minority "baseline"	0.28	0.25	0.69	0.62	0.64	0.57	0.59	0.56

(0.69/0.67 vs. 0.31), desired solutions (0.60/0.62 vs. 0.38), solution format (0.60/0.61 vs. 0.36), steps to perform tasks (0.56/0.58 vs. 0.43), resources (0.56/0.60 vs. 0.41), and acceptance criteria (0.57/0.55 vs. 0.44), while they failed to outperform the majority baselines overall unclarity (0.69/0.71 vs. 0.72) and difficult wording (0.71/0.71 vs. 0.75). Besides, the hypothetical strategy of predicting the minority class would result in competitive outcomes for some flaws. Comparing the two BERT versions, we saw that the case-insensitive version (Bert-base-uncased) functions slightly better, delivering higher results in five cases. This finding implies that our choice to lowercase all words for the features was suitable.

Given that the training set was not massive, the feature-based classification models benefited from their focused analysis; they acquired higher test set accuracy compared to both BERT classifiers for some cases: unspecified desired solution (A_5 and $A_{\setminus 6}$ with 0.63 vs. Bert-base-uncased with 0.62), solution format ($A_{\setminus 2}$, $A_{\setminus 4}$, and $A_{\setminus 5}$ 0.62 vs. 0.61), steps to perform the task ($A_{\setminus 2}$ 0.62 vs. 0.58), required resources (A_1 and $A_{\setminus 2}$ 0.63 vs. 0.60), and acceptance criteria (A_3 0.63 vs. 0.57). For overall unclarity, the SVMs $A_{\setminus 2}$ and $A_{\setminus 5}$ functioned best with 0.74 and proved a learning success against the majority baseline in classifying the task descriptions based on defined clarity dimensions. Particularly the models without the length feature ($A_{\setminus 2}$) appeared strong in general (more on the features below).

Eventually, the results indicated that the clarity of having a *difficult wording* seemed hard to evaluate; none of our methods outperformed the majority baseline (0.75). A reason might be the diversity of potentially complicated words, making learning such words challenging. Nevertheless, the majority baseline result also revealed that this clarity flaw revealed a relatively high distribution imbalance.

RQ2. Impact of the Textual Properties of Descriptions on Task Clarity The separate feature type results in the top part of Table 4.3 (A_i) indicate that many of the examined textual properties are appropriate for assessing at least some of the clarity flaws. The *content* of task descriptions (measured in the form of TF-IDF) seems particularly significant, acquiring notably higher results than the other feature types for several clarity flaws, including for overall unclarity (0.72). The *style* of the descriptions functions best on *unspecified acceptance criteria* (0.63), perhaps due to the specific part-of-speech tags, and it is also an essential indicator for an *unspecified solution format in descriptions* (0.60). Similarly, the *readability* of task descriptions enables best to determine *unspecified desired solutions* (0.63), whereas the *flaw-specific* features and the *subjectivity* of descriptions play a vital role mainly in an ablation setup in the center part of Table 4.3 ($A_{\setminus i}$).

The insightful exception among the eight feature types is the *Length* (representing the number of words, digits, characters), which gains comparably low precision for all considered clarity flaws of descriptions. The accuracy of the *Length* in the ablation setting is also underlined, meaning that the best results overall are acquired when the length feature is excluded. This finding indicates that the description clarity is independent of their length—which contradicts the related assessment tasks, such as predicting Wikipedia article quality (Lipka and Stein, 2010) or argument quality (Wachsmuth and Werner, 2020).

4.5 Conclusions

Prior research in crowdsourcing has primarily focused on the quality of solutions provided by crowd workers. The presence of low-quality solutions is seen as a

major obstacle in realizing the full potential of crowdsourcing. This challenge arises from various factors involving workers, requesters, and platforms. In particular, poor task design can lead to worker dissatisfaction and frustration due to misaligned expectations and unjustified rejection of their work. Such issues can strain the relationship between requesters and workers, disrupting the overall dynamics of crowd work over time.

Creating clear and comprehensive task descriptions poses a significant challenge in crowdsourcing, especially for inexperienced requesters. Unclear or incomplete instructions can have a detrimental impact on the quality of workers' results, affecting their rewards and reputation. In this regard, we propose that natural language processing techniques can effectively address this challenge by identifying clarity flaws in task descriptions.

This chapter aimed to examine the extent to which defined clarity flaws in task descriptions can be automatically determined by leveraging natural language processing techniques (RQ1), and we have studied the textual properties of task descriptions that show task clarity flaws (RQ2). For that, we have specified seven clarity flaws from relevant literature, all affecting a task's overall clarity (i.e., how much a task description is unclear). Due to the lack of a valuable dataset for investigating task clarity assessment, we have expanded an available dataset with flaw annotations on this basis. To this end, we have requested crowd workers to annotate the specified clarity flaws in 1332 real-world task descriptions.

We have addressed RQ1 by estimating the effectiveness of two types of computational methods for clarity flaw assessment: transformer-based classifiers (using BERT) and linear SVM Classifiers with feature types, such as standard content and style features and flaw-specific characteristics. In light of RQ2, we conducted an individual features analysis employing the SVMs, giving insights into the effect of the textual properties of descriptions on the clarity flaws assessment. Regarding RQ1, we discovered that the accuracy of the BERT models varies from 0.55 to 0.71. The SVMs beat the BERT's performance, with results between 0.61 to 0.74 for the majority of clarity flaws. Both methods indicated learning success in almost all cases, excluding determining difficult wording. For RQ2, we observed that descriptions' content, style, and readability are shown to be significant textual properties for clarity. Combining the task flaw-specific characteristics with others is also beneficial for clarity assessment. Contrarily, the descriptions' length was not helpful in identifying the clarity flaws.

In Chapter 3, we conducted an extensive study to gather crowdsourcing challenges, with a particular focus on the workers' perspective. This approach allowed us to obtain a comprehensive view of the existing problems and their significance in crowdsourcing processes, providing valuable insights into various areas that require improvement.

An analysis of the findings from the previous study revealed that the most critical problem leading to numerous challenges for crowd workers in crowdsourcing processes is requesters' failure to design tasks, particularly in terms of providing clear instructions. This lack of clarity results in low-quality submissions by workers, known as the most significant problem from requesters' perspective in crowdsourcing. In this chapter, we employed a methodology to evaluate the effectiveness of computational approaches in assessing crowdsourcing task description clarity. The study findings indicate that computational techniques can indeed offer us the capability to automatically assess clarity with respect to defined flaws.

In the upcoming chapter, we will introduce a tool that we have developed to aid requesters in improving the clarity of their task descriptions before posting them on crowdsourcing platforms. Our goal is to contribute to the existing knowledge and enhance workers' understanding of tasks, ultimately leading to improved quality of results. By offering clearer instructions, we anticipate that workers will experience greater satisfaction and reputation in crowdsourcing processes.

Chapter 5

Automated Writing Assistance

In this chapter, we present our computational approach and the novel tool called "ClarifyIt," both of which represent significant contributions to the advancement of methods aimed at enhancing task clarity in crowdsourcing marketplaces. We outline the approach we adopt to develop and evaluate the ClarifyIt tool, with the primary objective of addressing the prominent issue of unclear task descriptions that has been identified as a major challenge in crowdsourcing. The ultimate goal of ClarifyIt is to assist requesters in improving the clarity of their task descriptions, thereby increasing the likelihood of obtaining higher-quality results from crowd workers. This chapter delves into the methodology employed for the development and evaluation of the ClarifyIt tool, providing a comprehensive analysis of the evaluation results.

As discussed in Chapter 4, the quality of results the crowd provides has been the focus of extensive prior research on crowdsourcing (Kittur et al., 2013). Suboptimal results are a tremendous challenge in exploiting the full potential of crowdsourcing (Weld et al., 2015). Of the numerous factors affecting crowd work quality, the significance of ambiguous task design has been emphasized (Manam and Quinn, 2018). Therefore, writing clear task descriptions is vital for a practical task design. For instance, Table 5.1 displays two variations of the same task description with a title and a body of instructions, giving obvious differences concerning the define clarity dimensions (introduced as *clarity flaws* in Section 4.2). The instructions quality instantly influences the workers' perception, approval rate and eventually, their trust, satisfaction, and the final quality of results (Wu and Quinn, 2017). Although a clear task design is essential for crowdsourcing processes, ambiguous task descriptions have yet been emphasized as a constant challenge (Khanna et al., 2010, Chandler et al., 2013, Gadiraju et al., 2017, Gaikwad et al., 2017, Wu and Quinn, 2017, Nouri et al., 2020).

This problem lies in a dual complexity: First, requesters should sufficiently explain all required details for completing a task; however, this is usually difficult without broad crowdsourcing experience, particularly for micro-tasks having an expansive range of potential crowd workers, who are from diverse cultures, with

TABLE 5.1: Example of a pair of task descriptions from crowdsourcing, where the left description represents the initial version created by a ClarifyIt user, and the right description showcases the most improved version based on the tool's clarity score (Nouri et al., 2023).

	First Version of Task Description	**Best Version of Task Description**
Title	Creation of new pieces of writing	Creation of new pieces of writing pieces of text on arbitrary topics
Body	You will be responsible for writing	You will be responsible for writing pieces of text on a variety of different topics. These topics may be selected randomly. You will be provided with a topic to write about as well as a minimum necessary word count for the piece. You should provide the writing in a typed format, which will then be submitted by email. Once your writing has been assessed, you will be compensated if you have met the criteria. If you do not meet the criteria, you will be given feedback and a further opportunity to adjust your writing and resubmit.

various skills, and different educational backgrounds (Difallah et al., 2018). Second, creating clear and comprehensible instructions is naturally complicated due to both the subjective requesters ' perspective and the generally inherent ambiguity of natural language. Hence, crowd workers may interpret the instructions differently (Franklin et al., 2011). Arguably, creating an automated tool that helps task requesters write clear and complete instructions can support addressing the dual challenge of describing all necessary detailed information in precise phrasing. To the best of our knowledge, such a tool has not been publicized, probably due to the shortage of useful computational models that can assess the clarity of task descriptions.

In this chapter, we advance the state of the art in supporting crowdsourcing task design, as explained in Section 2.3, by introducing our computational method for automatically rating clarity flaws in task descriptions. Our approach leverages natural language processing techniques to facilitate the creation of a tool named *ClarifyIt*. This tool aids requesters to progressively enhance the clarity and quality of their task descriptions prior to publishing them on a crowdsourcing platform. Ultimately, this approach enables workers to confidently accept tasks with more detailed and transparent descriptions.

The development of the ClarifyIt tool involves creating computational models using essential components derived from the assessment study discussed in Chapter 4. Specifically, these key components encompass the defined task clarity flaws,

and the feature types designed for linear feature-based modeling techniques, which were explained in depth in Section 4.2; and the annotated dataset for crowdsourcing task description clarity, elaborated in Section 4.3.

In the subsequent sections, in Section 5.1, we will present an outline of the approach employed to accomplish the central goal of this chapter. We will then proceed to a more comprehensive exploration of the computational models harnessed by ClarifyIt in Section 5.2. The design and implementation of ClarifyIt's workflow will be elucidated in Section 5.3, followed by the evaluation process, which encompasses two distinct user studies: one involving requesters in Section 5.4, and the other involving workers in Section 5.5. Finally, we will provide a concise summary of our work in Section 5.6.

5.1 Approach

Figure 5.1 provides an overview of the method we employ to develop and assess our computational approach, aimed at assisting task requesters in crafting instructions that are not only more precise but also contain all the requisite detailed information. Our approach comprises two primary phases, specifically the *Development* and *Evaluation* phases, which are elaborated upon below:

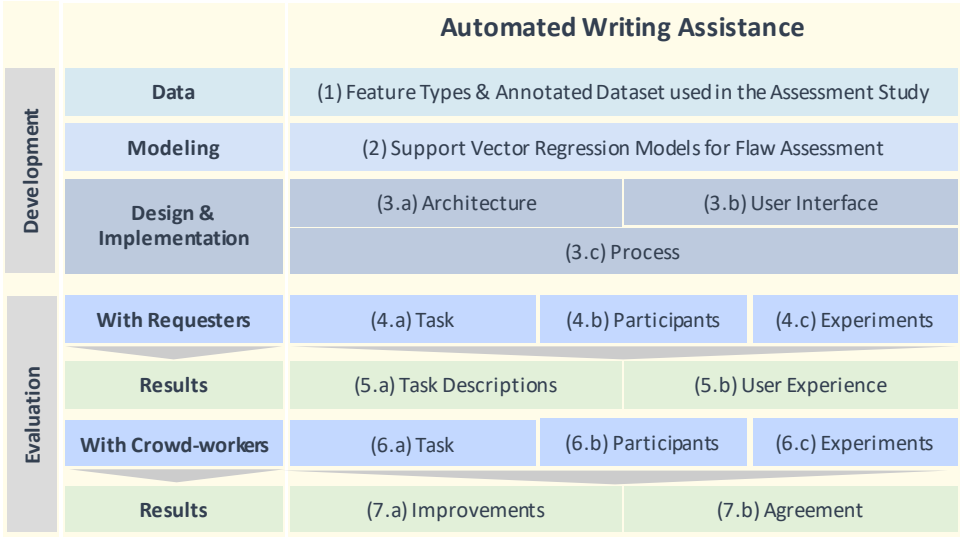


FIGURE 5.1: Overview of the approach used in the development and evaluation of our solution for unclear crowdsourcing task descriptions.

Development phase In this phase, we construct and put into operation the computational models for scoring the clarity of task descriptions based on the predefined clarity issues. These models are instrumental in the creation of a web-based

tool that enables requesters to compose task descriptions and enhance their clarity through iterative improvements.

- As preliminary steps for the modeling phase (Fig. 5.1(1)), we employ the feature types discussed in the study of computational methods for assessing task description clarity (Section 4.2). We utilize these features in conjunction with the annotated dataset (Section 4.3) pertaining to the defined clarity flaws in task descriptions.

- For modeling, we deviate from the previous study in Chapter 4, where we built binary classifiers, we employ Support Vector Regression (SVR) models (Fig. 5.1(2)). The SVR models enable our tool to assign an automated clarity score to task descriptions, going beyond a binary assessment of whether they are clear or unclear. Section 5.2 details the development of the computational models.

- Given the regression models for clarity flaw prediction, we design the ClarifyIt tool's architecture, user interface, and process (Fig. 5.1(3)) through which the requesters can iteratively improve their task descriptions. Section 5.3 elaborates on the creation of the ClarifyIt tool.

Evaluation phase In this phase, we aim to address the following two research questions:

> **(RQ1)** Based on the requesters' assessments, how helpful is the tool to identify and improve the clarity flaws in task descriptions?
>
> **(RQ2)** Based on the crowd workers' assessments, how effectively does the tool support creating clearer task descriptions in terms of completeness and comprehensiveness?

Specifically, we devise two consecutive user studies:

- First, we assess the helpfulness of our approach for requesters (Fig. 5.1(4)). In this phase, we formulate the evaluation task, outline the criteria for selecting participants, execute the experiment, and subsequently gather the outcomes of the user study (Fig. 5.1(5)). These outcomes include task description pairs produced by the study participants, as well as their feedback and comments related to their experience with our tool. The results indicate to what degree our tool is helpful for requesters (i.e., the only users of our tool) in improving their description clarity (RQ1). Section 5.4 comprehensively explains the evaluation study conducted with requesters.

- Based on the findings from the user study involving requesters, we proceed to conduct a second user study involving crowd workers (Fig. 5.1(6)). This

study is aimed at assessing the impact of our tool, ClarifyIt, on enhancing the clarity of task descriptions from the perspective of workers. Concretely, we incorporate a collection of task descriptions generated in the preceding user study into our task design. Similar to the prior study, we establish participant selection criteria, conduct the survey with crowd workers, and collect the results submitted by the workers. These results (Fig. 5.1(7)) offer valuable insights into the extent of improvement in the task descriptions facilitated by our tool. Additionally, they provide an assessment of the participants' agreement on the effectiveness of our tool in aiding requesters to enhance the clarity of their initial task descriptions (RQ2). Section 5.5 details the evaluation study with crowd workers.

By employing this approach, our aim is to make a valuable contribution to resolving the issue of unclear task descriptions in crowdsourcing. Our focus is on developing computational models and a tool that enables requesters to create clearer task descriptions effectively. In summary, the main contributions of this chapter are as follows:

- Computational models: we develop feature-based regression models to evaluate the degree of clarity in textual task descriptions, taking into account eight specific clarity flaws in crowdsourcing task descriptions.
- An automated writing assistance tool: we build an assistance tool to support requesters in enhancing the clarity of their task descriptions before posting them on a crowdsourcing platform, using an iterative approach.
- Extensive empirical insights into effectiveness of an automated writing assistance: we design and run two user studies to assess the effectiveness of an automated assistance in enhancing the clarity of task descriptions. enhancing the clarity of task descriptions. This assessment considers the perspectives of both the individuals requesting the tasks and the crowd workers performing them.

In the upcoming sections, we will outline the step-by-step approach we have designed to address the issue of incomplete and unclear task instructions in crowdsourcing platforms. The subsequent section delves into the creation of computational models using the elements derived from the evaluation study detailed in Chapter 4.

5.2 Computational Models

Our main objective is to design an automated writing assistance to help crowdsourcing task requesters identify their description clarity flaws and improve them iteratively. To this end, we require to build computational models that facilitate

developing such an automated assistance. In this section, we explain the computational models we developed to automatically score the clarity level of descriptions with respect to the flaws. We relied on the findings of the study in Chapter 4 that evaluated the usefulness of two distinct types of models in clarity flaw classification comparing a feature-based support vector machine (SVM) (Joachims, 1998) and a transformer-based BERT (Devlin et al., 2019). BERT models showed no consistent enhancements in clarity flaw detection; however, the SVM functioned more effectively in this specific use case and much more efficiently. Thus, we used the feature-based methods here, but we developed models that assign numerical values to quantify clarity flaws, enabling us to offer graded feedback on the level of clarity in task descriptions.

Our tool incorporates computational models to assess the degree to which each clarity flaw exists in a given task description. Contrary to the study in Chapter 4 that built binary classifiers, we utilized supervised *regression* models to acquire numerical scores denoting the unclarity degree for each flaw in descriptions. Since we seek to investigate how to support requesters to improve their task description clarity rather than studying the best regression approach, we decided to rely on the previous study findings in Chapter 4 and used support vector regression (SVR), known as the best techniques for feature-based regression. Given the full corpus from Section 4.3, we trained one distinct SVR model for each clarity flaw, obtaining eight independent SVR models in total.

To optimize the performance of each regressor, we selected the ideal set of features for the regressor by employing the *SelectKBest* class from *Scikit-learn*[1] which ranks the efficacy of all features and keeps only the k highest useful features to predict the unclarity level of a given description for a given k. We tested *SelectKBest* on SVR with 20 cost hyperparameters (in the range: 2^i for $-10 \leq i \leq 10$) and 15 separate values of k (in the range: $100 * i$ for $1 \leq i \leq 15$) for each clarity flaw regressor. In total, we obtained 300 distinct models for each clarity flaw and chose the hyperparameters of the best-performing model. We calculated the mean squared error employing 5-fold cross-validation to find the best-performing regressor for each clarity flaw. Finally, we trained each regressor using the corresponding feature sets and optimized hyperparameters for each clarity flaw.

In order to generate the confidence score for each dimension's prediction, we computed the standard deviation of the predictions from the top three best-performing pre-trained models, scored on a scale from 0 to 100.

5.3 ClarifyIt: A Tool to Write Task Descriptions

Employing the regression models introduced in the previous section, we developed *ClarifyIt* ('It' refers to both the Iterative process and the task description) to help

[1] scikit-learn library's link: `https://scikit-learn.org/stable/`

crowdsourcing task requesters write task instructions and iteratively improve their clarity. This section represents the tool's user interface and its designed process.

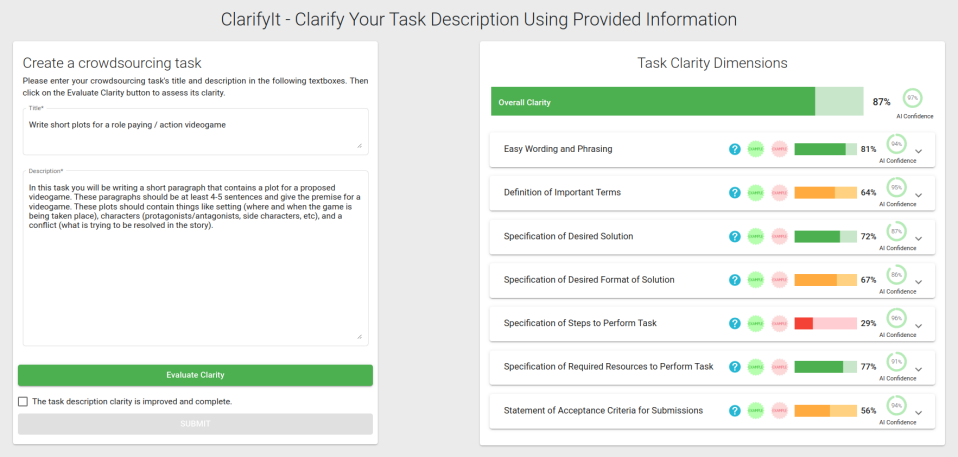


FIGURE 5.2: An image of the user interface for our writing assistance tool, *ClarifyIt*: The left side is where the requester inputs the title and task description. Upon clicking "Evaluate Clarity," the tool assesses the clarity of the task description and displays clarity scores for different flaws on the right side. The requester can continually refine the description by clicking the "Evaluate Clarity" button and making improvements until the description achieves clarity (Nouri et al., 2023).

Architecture ClarifyIt, as a web-based assistance tool, uses a three-layered architecture. The architecture has (a) the *presentation layer* (frontend) provisioning a user interface that crowdsourcing requesters use to interact with the system; (b) the *application layer* (backend) governing the calculation of clarity scores as well as logging and alike; and (c) the *data layer*, that supplies the pre-trained models and logs. HTML/CSS and Angular are used to implement the frontend, and Python to implement the backend.

User Interface Figure 5.2 demonstrates our tool's user interface, which has two main sections: (a) the *input section* on the left side of the User Interface (UI) is where requesters can write a task instruction and let the system evaluate its clarity; and (b) the *evaluation section* on the right side of the UI displays the task description clarity dimensions, their corresponding predicted scores, and the confidence scores. For each clarity dimension, the evaluation section also contains a concise description, a static example of the good and bad task description, meaning with

and without that clarity flaw. The user should click on the respective icons [2], , and [3] to view the description and the good and bad examples.

Process The requester can use our tool to draft a description (containing a title and a body) from scratch or paste an already-written version outside the tool. When the user clicks on the *Evaluate Clarity* button, the task description is sent from the presentation layer to the application layer for the clarity score predictions. The task description is passed through all feature-type modules in the application layer to compute their corresponding feature values. The pre-trained regression models in the data layer are fed by the feature values. Then, one score for each clarity dimension, in terms of a percentage value, is calculated by scaling the score that the corresponding model predicts. Eventually, the predicted and confidence scores are shown on the presentation layer. Ultimately, the requester can then consider the predictions to enhance the description's clarity iteratively by redoing the illustrated process in the tool.

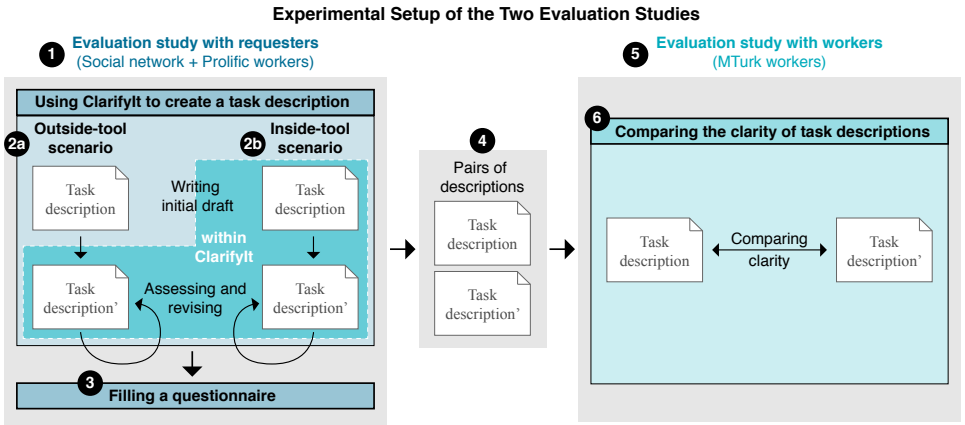


FIGURE 5.3: An overview of the experimental setup of the study that examines the helpfulness and effectiveness of ClarifyIt with task requesters (the gray box on the left side) in writing clear task descriptions and with workers (the gray box on the right side) in better understanding the task descriptions (Nouri et al., 2023).

5.4 Evaluation with Task Requesters

Before evaluating the quality of the written description utilizing our tool, we designed an experiment process by studying the impact of ClarifyIt on requesters and workers pertaining to task descriptions. Figure 5.3 illustrates the step-wised

[2]Question Mark (`https://icons8.com/icon/80684/question-mark`) icon by https://icons8.com

[3]Example (`https://icons8.com/icon/kP5VLEsdwqY8/example`) icon by https://icons8.com

experimental setup. Initially, we assess the tool's effectiveness with requesters (cf. ❶) who create crowdsourcing task descriptions using two distinct scenarios: the outside-tool scenario (cf. 2a) and the inside-tool scenario (cf. 2b). Subsequently, we collect feedback from them through a questionnaire (cf. ❸) regarding their experience of using the tool.

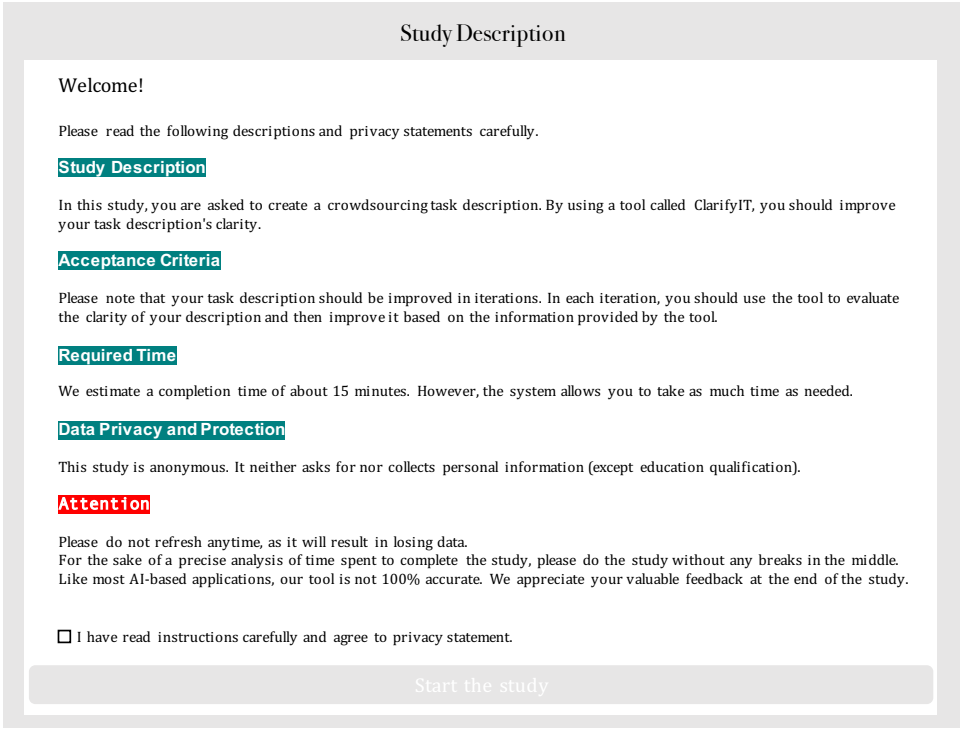


FIGURE 5.4: The task description of the user study with requesters shown at the first page of the study when requesters accepted to participate in the study.

In this user study, we evaluate the requesters' opinions on the tool's helpfulness in assisting them by providing the functionalities and necessary information to write a clear task description. We also study how effectively the tool enhances description clarity using the scores provided by the computational models. In the following, we describe the user study with requesters in detail, instructing them to compose a task instruction in an iterative manner using ClarifyIt and then answer a questionnaire about their experience. Figure 5.4 shows the task instruction we created for the user study with requesters, clarifying the goal of the task, acceptance criteria, required time and the data privacy.

TABLE 5.2: A list of designed scenarios for well-known micro-task types which were given to requesters in the evaluation study each of which aims to provide a target requester with an idea of creating a task descriptions.

#	Task Type	Scenario
1	Content Access (CA)	Imagine a situation where you need feedback for a piece of work, result of a search on the web, like/vote for an arbitrary post on social media. Please write down a task description explaining your task to crowd workers.
2	Content Creation (CC)	Imagine a situation where you need transcriptions of an audio file, a piece of text written about an arbitrary topic, or title/description for an arbitrary object. Please write down a task description explaining your task to crowd workers.
3	Interpretation & Analysis (IA)	Imagine a situation where you have an entity like a set of images, objects, audio files, or similar to be tagged or rated according to some conditions. Please write down a task description explaining your task to crowd workers.
4	Information Finding (IF)	Imagine a situation where you need information about an object, person, company, brand, or similar. Please write down a task description explaining your task to crowd workers.
5	Survey (SU)	Imagine a situation where you have created a questionnaire which includes questions about a specific subject. Please write down a task description explaining your task to crowd workers.
6	Validation & Verification (VV)	Imagine a situation where you like to assess the quality or correctness of an entity like a website, a piece of text, or similar. Please write down a task description explaining your task to crowd workers.

Experimental Setup

We set up the user study with requesters as follows:

Task Table 5.2 introduces the six designed scenarios for six crowdsourcing task types to avoid causing the participants to deal with an unknown domain. We defined and randomly assigned each scenario to participants, asking them to imagine themselves as the task scenario's requester, write down a description, and enhance its clarity using ClarifyIt.

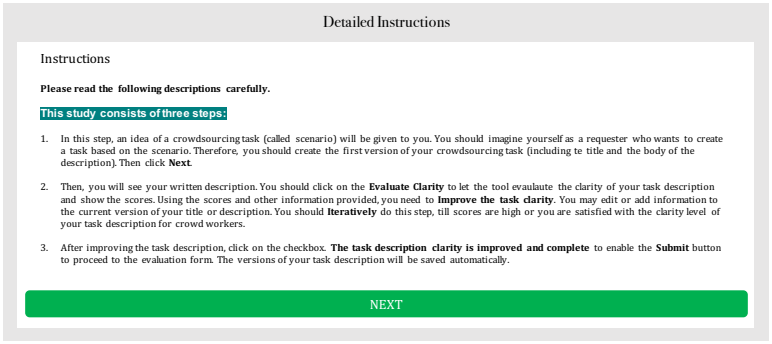
More specifically, Figure 5.4 shows the landing page of the user study with requesters, where we provided information such as a brief study description, the acceptance criteria, the required time to complete the study, data privacy policy, and notable details to avoid the task interface from crashing in the middle of the study. Finally, the participants state that the information has been carefully read, and start the study.

Participants As task requesters in the user study, we invited (a) researchers from our community and (b) crowd workers on *Prolific* (Fig. 5.3 ❶). 14 researchers from our social network and 108 from prolific (i.e., 122 participants) completed our study. We required the prolific workers to have an approval rate higher than 95%, at least 100 previous task submissions, and speak English as their first language. Deploying to a pilot study, we estimated 15 minutes to complete the task and paid £2.50 (meaning £10.0/hour). The researchers from our network did their work for free.

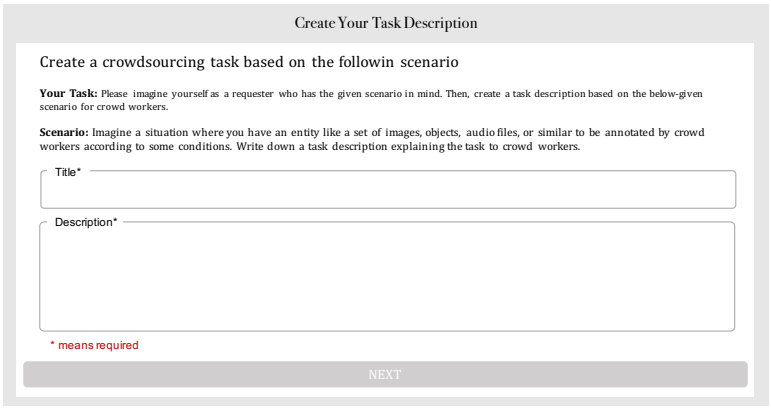
Experiments After starting the study from the landing page, the participants were randomly assigned to one of two study settings: outside-tool and inside-tool scenarios. These settings were designed to investigate how effectively the information provided by our tool enhances the clarity of initial task descriptions. We examine whether the cause of unclear task description lies in requesters' limited knowledge and experience regarding the essential information required for workers to understand the expected quality of their submissions in a clear task description. Specifically, our aim was to examine whether only displaying the clarity dimension at the outset, before providing the clarity scores, affects requesters' abilities, particularly beginners, to include the necessary details in their instructions. In the following, we explain the two study settings:

Outside-tool scenario Figure 5.5 shows this scenario's steps, corresponding instructions and activities. In this setting, participants view the instructions in detail (Fig.5.5(a)) and create an initial task description having the given scenario in mind (Fig.5.5(b)) before entering the tool and viewing the clarity dimensions on the user interface. After entering the tool, the initial task descriptions are automatically copied into the tool (Fig.5.5(c)), and participants start evaluating the clarity and refining them (Fig. 5.3 ❷a).

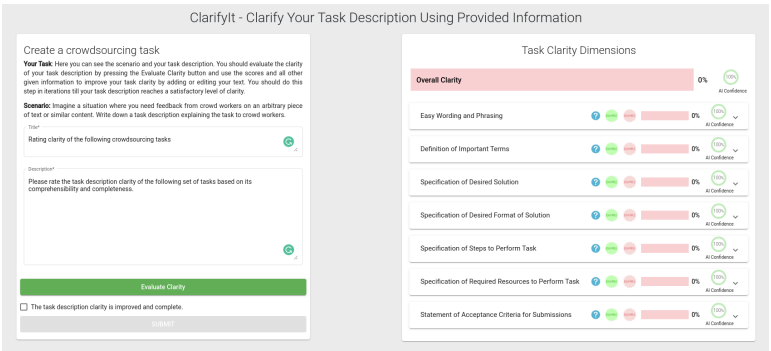
Inside-tool scenario Figure 5.6 shows this scenario's steps, corresponding instructions, and activities. In this setting, participants view the instructions in detail (Fig.5.6(a)) and then enter the tool (Fig.5.6(b)). They read the scenario there and then require to create, evaluate, and improve the description in the tool (Fig. 5.3 ❷b).



(a) Detailed instruction of the outside-scenario study.

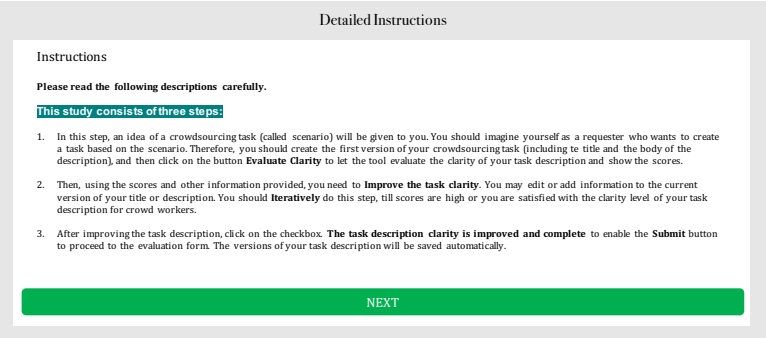


(b) The place where the requesters write the initial task description before entering the tool.

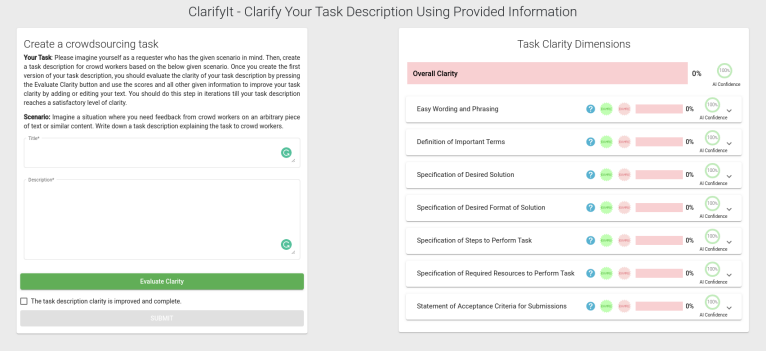


(c) After writing the initial description, requesters enter the tool where the initial version is automatically copied.

FIGURE 5.5: The outside-tool scenario's steps of the evaluation study with requesters where initial description are written before entering the tool.



(a) Detailed instruction of the inside-scenario study given to requesters before writing the task description.



(b) The place where the requesters view the task scenario according to which they write the initial task description, while viewing the clarity dimension on the tool.

FIGURE 5.6: The inside-tool scenario's steps of the evaluation study with requesters where participants first enter the tool and can write their initial task description using the information provided on the tool.

Overall, participants were to create the initial task description and then iteratively evaluate and improve the description's clarity considering the clarity dimension scores and other information shown by the tool. The iteration of evaluation and clarity improvement may continue until the participants acknowledge that the description reaches good clarity—according to the tool's scores or their judgment.

Figure 5.7 shows the questionnaire the participants were to answer after finalizing the task description (Fig. 5.3 ❸). The questionnaire contained 17 questions about the user experience with crowdsourcing in general (such as length and platforms) and ClarifyIt in the study. Eventually, requesters could submit suggestions to improve our tool's usability. We discuss the result of the user study with requesters in the following.

FIGURE 5.7: The questionnaire designed to evaluate the helpfulness of our tool. The participants of the user study with requesters filled in this questionnaire at the last step of the study.

Results

Altogether, 122 participants having up to 13 years of experience posting tasks on crowdsourcing platforms finished our study. 107 participants were *novice* with no prior experience and the remaining 15 were *experienced* requesters. They mostly

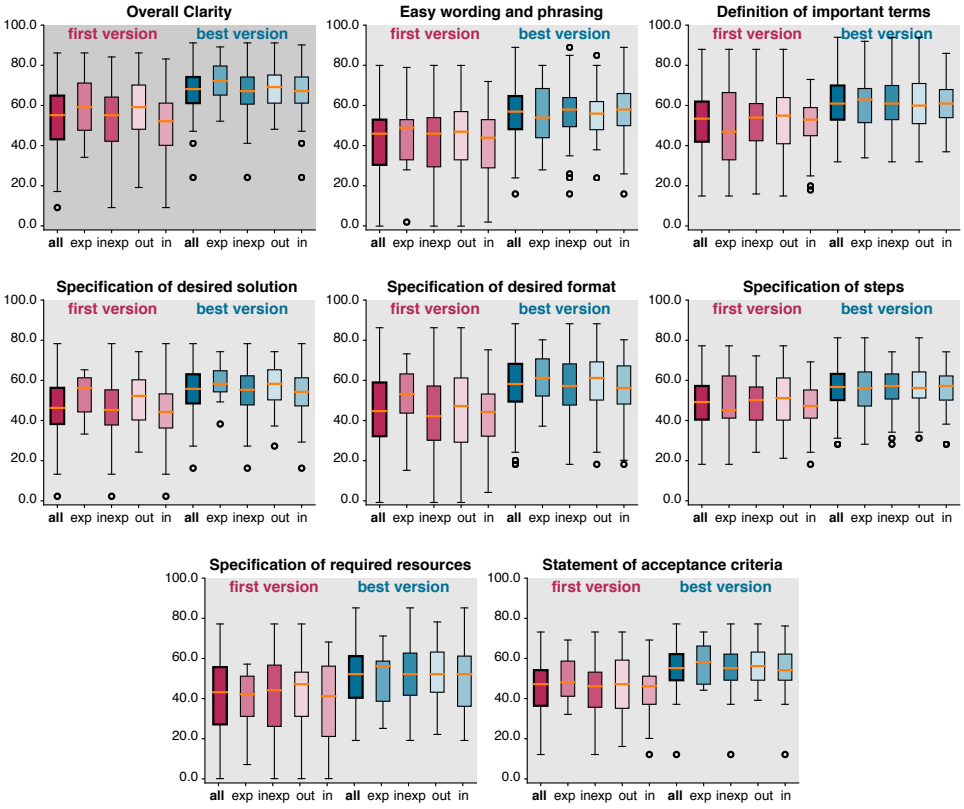


FIGURE 5.8: User study with task requesters: The scores of all eight considered clarity dimensions for the *first version* of the requesters' task description as well the *best version*, manually created and automatically scored using our tool, ClarifyIt: The box-and-whiskers plots show the results of *all* participants, *exp*erienced vs. *inexp*erienced participants as well as *out*side-tool scenario vs. *in*side-tool scenario participants. In all cases, the scores improved notably from the first to the best version (Nouri et al., 2023).

knew Amazon Mechanical Turk, Toloka, and Prolific. The outside-tool scenario had 57 successful submissions, and the inside-tool scenario had 65 submissions.

Although not given in the study instructions, 24 participants assigned to the outside-tool scenario (42%) modified their task description directly after viewing the clarity dimensions shown on the tool before checking the clarity. Therefore, the average overall clarity improved by eight percentage points, indicating that the information on the tool is effective from the beginning for writing a clear task description.

Task Descriptions Figure 5.8 indicates that, on average, participants enhanced their task descriptions clarity utilizing ClarifyIt. The best-scored version of task descriptions on all eight clarity dimensions remarkably improves over the initial version. All dissimilarities are noteworthy at $p < 0.01$ according to a paired t-test. Moreover, they all show a medium or extensive effect size. For example, the

difference between the best-scored (M = 64.61, SD = 10.87) and the initial version (M = 50.75, SD = 16.20) of the overall clarity dimension has a *Cohen's d* value of 1.01, showing a large effect size.

Figure 5.8 likewise shows that the inexperienced participants more clearly define important terms and the steps to complete the task in the first version. Nevertheless, the initial version's overall clarity, wording, desired solution, and format written by experienced participants are clearer according to our tool. This observation suggests that experienced requesters occasionally neglect clarifying new terms that may be unclear for workers from outside the domain. Furthermore, completing the task may sound vital to the task requesters. Thus, they overlook providing essential information on how the crowd workers should complete the task and submit their work.

We also observe that the first version of the descriptions written through the inside-tool scenario has no higher clarity score than the initial descriptions created through the outside-tool scenario. We can interpret that the general knowledge of clarity aspects of task descriptions does not influence the descriptions' clarity. Yet, the score of each clarity aspect for a given description can help the writer improve the clarity. Altogether, the results also show that the clarity of the best-scored descriptions improved using ClarifyIt in all cases.

Although participants mostly enhanced the clarity of their task description iteratively, 24% of the iterations reduced the dimension scores, implying that modifications to the descriptions through those iterations decreased their clarity (as evaluated by the tool). Additionally, we discovered that the overall clarity score of the final version of descriptions written by 42 participants (34%) is lower than the best score version of those descriptions, raising the necessity for having an undo functionality in ClarifyIt so that enables requesters to revert to the previous version of the description when the score falls. A few participants suggested the same in the comments. For instance, one requester mentioned:

> R1: *"I would like to see the previous score in order to get an idea about the archived improvements."*

User Experience Regarding the research question RQ1, Figure 5.9 shows the requesters' answers to the questionnaire in Figure 5.7. The answers indicate that the most helpful feature of our tool is the *general functionalities* with almost 65% being positive about it (13.11% very high, 51.64% for high) and just 12% negative (0.82% very low, 11.48% low). Additionally, 62% of the participants voted for the helpfulness of the information about the *clarity metrics*, and 60% for the *characterization of the clarity dimensions*. Participants also provided positive feedback about the overall usefulness of ClarifyIt in the open-ended comments. For instance, one requester explained the helpfulness of clarity dimensions as follows:

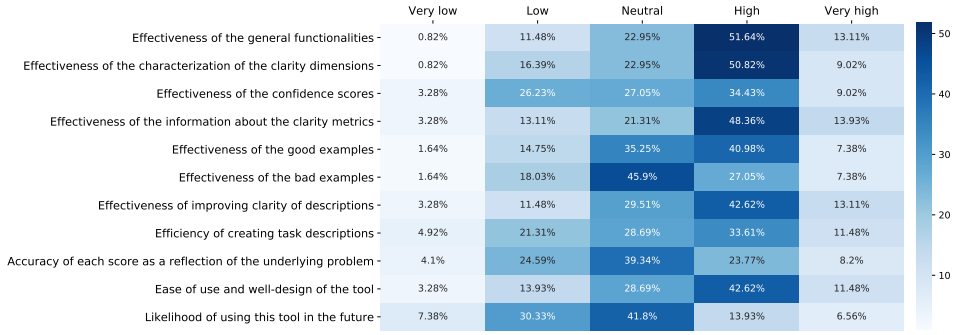


FIGURE 5.9: User study with task requesters: Distribution of the scores for the questions in the questionnaire about the experience with our tool, ClarifyIt. For most questions, most requesters saw the tool's effectiveness as high (Nouri et al., 2023).

> R2: *"The tool was extremely helpful. I lacked imagination in creating the task specifics however the metrics were genuinely helpful in clarifying what workers would need to know."*

Moreover, 56% of the requesters expressed that the tool is *easy to use and well-designed*, and 54% believed that it is very helpful in *improving the task descriptions clarity* . One requester wrote:

> R3: *"I could see through each of my edits how I was making the description clearer and easier to understand."*

The *good examples* given for clarity dimensions and the scores' accuracy for the unclarity assessment are relatively helpful, with just 16% and 38% of opposing opinions, respectively. The latter is probably because of the distributional change between the training's task descriptions and the participants'. For instance, one requester noted that our tool failed to identify the task description content corresponding to some dimensions:

> R4: *"Naming the categories helped by making it clear which aspects should be present in the description. Unfortunately, however, the AI did not recognize when I included the aspects I had previously forgotten in the description."*

The *bad examples* for clarity dimensions were neither detected as useful nor ineffective with 46% of neutral votes. We also learned that participants gave various answers to the *efficiency* of writing task descriptions with clarity, with 29% negative, 39% neutral, and 32% positive votes.

In total, 34% of the participants expected modification suggestions by the tool for improving the description clarity. Three of them are listed below:

> R5: *"[It] should tell you which parts to improve"*

> R6: *"It does not provide suggestions as to where it could be improved for example where you may benefit from adding a comma."*
>
> R7: *"[It] would be more helpful if it could show suggestions of better wording to improve it."*

42% of participants could not predict whether they will use our tool in the future, and 38% mentioned that they would likely not pursue help to form their task descriptions. Improvements in good or bad examples, in models' prediction quality, and dynamic task-related suggestions for clarity modifications can improve the tool's popularity and effectiveness in assisting requesters. They are potentially the research ideas in future work. More examples of the participants' comments on the tool are provided in Appendix A, Section A.3.1.

5.5 Evaluation with Crowd Workers

Given the findings from the evaluation study with requesters, we conducted a second evaluation study with crowd workers to examine the effectiveness of the tool, ClarifyIt, in enhancing the clarity of crowdsourcing task description from the workers' viewpoint (Fig. 5.3 ❺). In particular, we requested the participants compare the clarity of the first version and the best version of the descriptions the requesters wrote (Fig. 5.3 ❻) regarding the eight clarity dimensions to evaluate whether requesters successfully created clearer task descriptions utilizing the tool.

Experimental Setup

We set up the user study with crowd workers as follows:

Task A random sample of 100 pairs of task descriptions, generated during the study with requesters, was selected (Fig. 5.3 ❹). Each pair consisted of the initial version written by the requester and the best version, which had been iteratively modified based on the overall clarity score computed by the respective model. This set of task descriptions (examples in Appendix A, Section A.3.2) was used in the task design of the user study with workers to assess the tool's effectiveness in assisting requesters in enhancing their task descriptions.

Participants Since the actual task descriptions created for the assessment study (detailed description in 4.3) were initially posted on Amazon Mechanical Turk (MTurk) and likewise were annotated for clarity flaws by workers on MTurk, we planned to engage MTurk workers to compare the clarity of the given task description pairs.

For language proficiency reasons, we set our participant selection filter to only workers from Canada, the UK, the US, Ireland, New Zealand, Australia, and South Africa. Besides, participants were required to have at least 10,000 approved submissions on MTurk, and an approval rate of the lowest 98%. Aligned with our funding limitations, we hired seven workers to judge the clarity modification of each task description pair and rewarded each voter USD 1.25 for an estimated six-minute work. While more participants would additionally improve statistical reliability, seven opinions seem sufficient to determine general tendencies. To achieve higher results quality, we designed two attention checks introduced below to evaluate the quality of votes.

[DATA PRIVACY] The results are anonymous, i.e. all answers are recorded and stored anonymously. Your participation is voluntary. There are no disadvantages if you do not participate. Your information will be treated in strict confidence.

Instructions:

In this HIT, four pairs of task descriptions are given. You should compare the clarity of each pair of task descriptions which were created and improved by requesters using **ClarifyIt** tool which is developed to help requesters to make task descriptions clearer for crowd workers.

Each pair of task descriptions consist of a title and a body, and is followed by ten statements each of which compares different clarity aspects of the given task descriptions.

To increase the quality of your answers, please imagine yourself as a worker who wants to complete the task and receive the payment.

For statements 1 to 10, please carefully express the degree to which you agree with each statement for the given pair.

FIGURE 5.10: The summary of our comparison task description on the MTurk platform was shown to crowd workers when they clicked on the task in the list on the dashboard to view the details of the task description before the final task acceptance decision.

Experiments The task description of our user study with crowd workers was similar to the description shown in Figure 4.4 except for the payment condition, and was shown in the MTurk dashboard. Clicking on our task description in the dashboard, the study were shown to workers with a more specific instructions shown in Figure 5.10 followed by comparison blocks. Figure 5.11 and 5.12 show two out of four comparison blocks of a task example of the user study with crowd workers posted on MTurk. Each block shows one task description pair (i.e., Task Description 1 and Task Description 2) followed by the eight statements designed to compare the clarity of Task Description 1 and 2 regarding the respective dimension. We asked workers to choose to what degree they agreed with the statement on a 5-point Likert scale from *strongly disagree* to *strongly agree*.

Comparison #1

Please read and compare the two following task descriptions thoroughly and then, by carefully considering the instructions mentioned above, express the degree to which you agree with each statement for your comparison.

Task description 1

Title: Creation of new pieces of writing

Body: You will be responsible for writing pieces of text on arbitrary topics

Task description 2

Title: Creation of new pieces of writing

Body: You will be responsible for writing pieces of text on a variety of different topics. These topics may be selected randomly. You will be provided with a topic to write about as well as a minimum necessary word count for the piece. You should provide the writing in a typed format, which will then be submitted by email. Once your writing has been assessed, you will be compensated if you have met the criteria. If you do not meet the criteria, you will be given feedback and a further opportunity to adjust your writing and resubmit.

In 'Task description 1' compared to "Task description 2':	Strongly disagree	Rather disagree	Partly disagree Partly agree	Rather agree	Strongly disagree
1. It is clearer how to complete the task and what the desired solution is.	○	○	○	○	○
2. The wording and phrasing is clearer and easier to understand.	○	○	○	○	○
3. The potentially important terms are better defined.	○	○	○	○	○
4. There are more words in 'Task description 1'.	○	○	○	○	○
5. The desired solution is better explained in more detail.	○	○	○	○	○
6. The format in which the solution should be submitted is better specified.	○	○	○	○	○
7. The steps to complete the task are better specified.	○	○	○	○	○
8. The resources that are required to complete the task are more completely specified (such as data, tools, links, or websites).	○	○	○	○	○
9. The acceptance criteria for a solution to the task are better specified.	○	○	○	○	○
10. The potentially important terms are not sufficiently defined.	○	○	○	○	○

FIGURE 5.11: An example of attention check version 1 with one out of four comparison tasks that workers on the MTurk platform were requested to complete and submit.

Comparison #2

Please read and compare the two following task descriptions thoroughly and then, by carefully considering the instructions mentioned above, express the degree to which you agree with each statement for your comparison.

Task description 1

Title: Write short plots for a role playing / action videogame

Body: In this task you will be writing a short paragraph that contains a plot for a proposed videogame. These paragraphs should be at least 4-5 sentences and give the premise for a videogame. These plots should contain things like setting(where and when the game is being taken place), characters (protagonists/antagonists, side characters, etc), and a conflict(what is trying to be resolved in the story).

Task description 2

Title: Write short plots for a role playing / action videogame

Body: In this task you will be writing a short paragraph that contains a plot for a proposed videogame. These paragraphs should be at least 4-5 sentences and give the premise for a videogame.

In 'Task description 1' compared to 'Task description 2':	Strongly disagree	Rather disagree	Partly disagree Partly agree	Rather agree	Strongly disagree
1. It is clearer how to complete the task and what the desired solution is.	○	○	○	○	○
2. The wording and phrasing is clearer and easier to understand.	○	○	○	○	○
3. The potentially important terms are better defined.	○	○	○	○	○
4. There are more words in 'Task description 1'.	○	○	○	○	○
5. The desired solution is better explained in more detail.	○	○	○	○	○
6. The format in which the solution should be submitted is better specified.	○	○	○	○	○
7. The steps to complete the task are better specified.	○	○	○	○	○
8. The resources that are required to complete the task are more completely specified (such as data, tools, links, or websites).	○	○	○	○	○
9. The acceptance criteria for a solution to the task are better specified.	○	○	○	○	○
10. It is clearer how to complete the task and what the desired solution is.	○	○	○	○	○

FIGURE 5.12: An example of attention check version 2 with one out of four comparison tasks that workers on the MTurk platform were requested to complete and submit.

For the first attention check, we designed two objective statements to determine whether the workers carefully read the statements. Given the four comparison blocks per task, we added one as the fourth statement of comparison blocks #1 (Fig. 5.11) and #3, and another in #2 (Fig. 5.12) and #4. If the statement was correct for a pair, *strongly agree* or *agree*, if false, *strongly disagree* or *disagree* votes passed the attention check. Only participants who passed the check for four comparisons were evaluated for the second attention check.

For the second attention check, we duplicated an arbitrary statement of each comparison block in the last statement to check whether the workers carefully gave their opinion. Figure 5.11 shows the repeating of the third statement for the tenth, and Figure 5.12 the first. The requirement for passing this attention check was consistent votes (i.e., (strongly) agree or (strongly) disagree) for the statement in both occurrences for each comparison block. Because of the subjective nature of repeated statements, we assumed that workers might slightly change their thought. Therefore, *neutral* opinion was also acceptable, passing the attention check and obtaining the payment for the submission.

Results

We obtained 700 submissions from 92 individual workers comparing the 100 pairs of task descriptions. To collect uniform votes, we automatically inverted the votes for those pairs where the initial version was given as Task Description 1; denoting workers voted against higher clarity in the best-scored version.

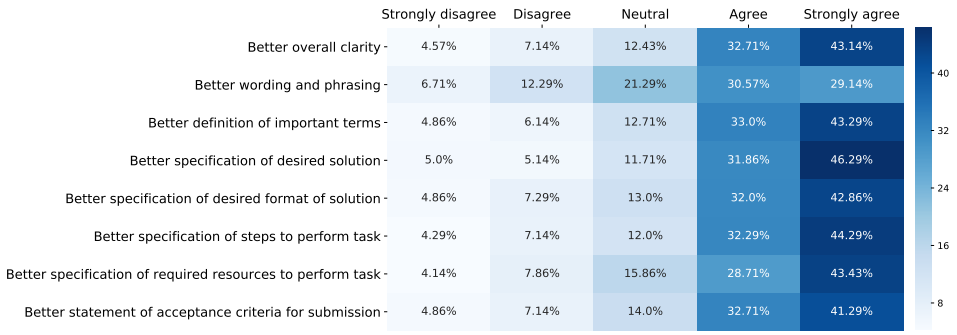


FIGURE 5.13: Evaluation with crowd workers: Distribution of the scores on improvements in terms of the eight clarity flaws of the best-scored versions of the 100 task descriptions over the initial version. Workers did not know which version is which one (Nouri et al., 2023).

Improvements Figure 5.13 shows the distribution of all workers' votes for improvements level in task descriptions' clarity. In light of RQ2, the descriptions' best-score version has higher clarity than their initial version concerning all clarity

dimensions. The most notable clarity modifications were in *specification of desired solution* (78%, 46.29% strongly agree and 31.86% agree), *specification of steps to perform task* (77%), *overall clarity* and *definition of important terms* (both 76%) dimensions based on all workers' votes. Additionally, 60%–75% of workers voted in favor of enhancements in other clarity dimensions. Nevertheless, improvements in *wording and phrasing* obtained the most significant proportion of negative votes (19%, 6.71% strongly disagree, and 12.29% disagree) as well as most neutral opinions (21%).

TABLE 5.3: Evaluation with crowd workers: (a) Average agreement of the seven workers on the 100 task description pairs, and (b) proportion of task descriptions whose clarity improved over the initial version by using our tool, ClarifyIt, according to the workers; both for each clarity dimension considering neutral votes either for (*case 1*) or against (*case 2*) improvements (Nouri et al., 2023).

		(a) Average agreement		**(b) Improved descriptions**	
#	**Clarity Dimensions**	**Case 1**	**Case 2**	**Case 1**	**Case 2**
1	Overall clarity	81%	89%	86%	98%
2	Wording and phrasing	69%	81%	68%	98%
3	Important terms	81%	90%	86%	98%
4	Desired solution for task	82%	91%	88%	98%
5	Desired format of solution	78%	89%	89%	97%
6	Steps to perform task	80%	89%	88%	98%
7	Required resources to perform task	79%	89%	82%	98%
8	Acceptance criteria for submission	78%	90%	88%	97%

Agreement We grouped the workers' votes into binary (positive or negative) labels, denoting votes for and against enhancements in the clarity of descriptions. The agreement of the seven voters for each description pair was calculated in two possible ways, considering the neutral votes as against (*Case 1*) and as in favor of (*Case 2*) clarity advancements in the descriptions' best-scored version.

Table 5.3(a) displays the average consensus for all clarity dimensions, and Table 5.3(b) the ratio of the clearer best-score version of task descriptions for the eight dimensions in both cases. In Case 1, the consensus among the voters ranges between 68% (for *better wording and phrasing*) and 82% (for *desired format better specified*), and the enhancement level in the clarity ranges from 68% and 89% (for the exact dimensions). In Case 2, the ranges shift to 81% (for *better wording and phrasing*), to 91% (for *desired format better specified*), and the enhancement level boosts to 97%–98% across the eight clarity dimensions. In light of RQ2, we con-

clude that the workers' view obviously suggests our tool's effect in writing clear task descriptions.

5.6 Conclusions

The challenge of unclear task instructions written by requesters constantly results in low-quality submissions by workers since it directly leads to misapprehension and misunderstanding of the tasks. Prior research determined such unclarity as one of the major problems lowering success in crowdsourcing.

In this thesis, we have investigated the impact of our computational models and assistance tool called *ClarifyIt* that we created to help requesters create and modify task descriptions iteratively until reaching sufficient clarity. The tool utilizes natural language processing methods to automatically predict the extent of eight common clarity flaws in task descriptions. The tool's workflow does not involve worker intervention, potentially increasing efficiency and effectiveness compared to prior solutions. In two user studies, we examined the effectiveness using the requesters' and workers' judgments. In the first study, requesters utilized our tool to write task instructions and improve clarity. We then asked the requesters to evaluate how much the tool helps improve clarity (RQ1). In the second study, the crowd workers compared the clarity of the initial and improved versions of task descriptions to judge whether our tool helped the requesters enhance their descriptions' clarity (RQ2).

Regarding RQ1, our findings suggest that, on average, participants improved the clarity of their task descriptions using ClarifyIt. The best-scored versions of the task descriptions showed significant enhancements across all eight clarity dimensions compared to the initial versions. These improvements were observed to have a medium or large effect size. Inexperienced participants tended to provide clearer definitions of important terms and task steps in their initial versions.

However, our tool identified that task descriptions written by experienced participants had higher overall clarity, better wording, desired solution, and format. This observation indicates that experienced requesters may sometimes overlook the importance of clarifying new terms that could be unclear to workers from different domains. Furthermore, they may not provide sufficient information on how workers should complete the task and submit their work. Altogether, the first study's outcomes suggest that the tool's primary utilities and provided information are beneficial. Besides, the requesters noticed the tool was well-designed and useful in specifying and enhancing a description's clarity.

Concerning RQ2, the crowd workers' votes determine that the eight clarity flaws of descriptions written using ClarifyIt considerably improved on average. However, the clarity of the wording and phrasing in the instructions is the most complicated dimension to anticipate computationally and, hence, to help requesters.

In the upcoming chapter, we will provide an overview of our entire approach to studying crowdsourcing problems. We will discuss the process of assessing the computational solution, followed by the development and evaluation of the solution presented in this thesis.

Chapter 6

Conclusion

Crowdsourcing marketplaces offer convenient access to a diverse range of human expertise, providing cost-effective solutions and services. This work model has the potential to benefit both requesters and workers. Requesters hope to receive high-quality submissions, while workers are driven by monetary or non-monetary rewards like building their reputation or enhancing their skills. Nonetheless, in reality, crowdsourcing models frequently face challenges in meeting the expectations of both requesters and workers.

The extensive body of literature on crowdsourcing has made significant contributions to our understanding of crowdsourcing, covering topics such as definitions, system design, pillars, applications, and processing methods. This research was part of the "Digital Future" program (introduced in Chapter 1), which aimed to deepen our understanding of crowdsourcing processes and develop technological methods that provide support to individuals, employers, and freelancers in their professional endeavors.

In the following sections, we provide an overview of the essential stages involved in the formation of this thesis. We discuss the motives behind the research questions tackled and highlight our main contributions to the state-of-the-art at each stage, emphasizing the significant findings acquired throughout this journey. Additionally, we explore prospective avenues for further research.

6.1 Contributions and Findings

In line with the primary goals of the "Digital Future" program, we first conducted a study to provide a broad overview of *Crowdsourcing Challenges* (detailed in Chapter 3) and their dominance from the crowd workers' viewpoint, due to the importance of their role in delivering excellent solutions to requesters. Leveraging the insights gained from this study, we focused on the challenge of unclear task descriptions provided by requesters which has been highlighted as the primary issue. Our main objective was to employ natural language processing techniques to develop an automated solution for helping requesters in creating task descriptions.

Therefore, our attention was directed towards *Task Clarity Assessment* (detailed in Chapter 4) to study whether clarity flaws in task descriptions can be automatically identified with trained models. In pursuit of this, we established the foundational components for our investigation, created the necessary dataset, constructed the models, and conducted a thorough analysis of the outcomes.

Building upon the insights gained from the assessment study, we advanced to develop an *Automated Writing Assistance* (detailed in Chapter 5). This involved crafting the necessary models using the components forged in the prior assessment study and implementing an interactive tool by applying these models. To measure the effectiveness of our solution, we conducted an evaluation that assessed how well our tool met our main objectives. The following will provide a summary of our primary discoveries and the contributions made at each stage.

Crowdsourcing Challenges

As discussed in details in Chapter 3, we aimed to obtain a thorough understanding of the obstacles that crowd workers face in the process. In particular, we concentrated on investigating the challenges faced by workers in the crowdsourcing process and identifying the most common issues.

Hence, we first conducted a thorough literature review to examine existing research and identify challenges discussed by scholars. We then performed an empirical data analysis by collecting and analyzing crowd workers' reviews on Turkopticon forum, where they exchange work-related experiences with their peers. Using topic models, we extracted challenges expressed by workers regarding complications, confusion, and unfairness experienced during their work. Our hypothesis was that these forum discussions contain valuable insights into practical problems faced by workers that may not have been extensively covered in the existing literature. Through our empirical analysis, we identified a set of challenges that were frequently discussed among workers in the forum.

We discovered that the primary issues workers complain about relate to task design. Requesters seem to undervalue the workload in terms of the time, effort, and fair compensation needed to complete a task. Additionally, they struggle with composing clear task instructions. Errors in their task environment implementation also lead workers to waste time on unsuccessful submissions. Some requesters are also reported to violate workers' privacy by not adhering to the terms-of-service of the crowdsourcing platform being used. Unexplained and unfair rejections are the predominant issue in task evaluation, and the literature extensively examined communication problems between requesters and workers, coupled with inadequate platform support.

Furthermore, our data analysis indicates that task design appears to be the primary source of issues from the workers' perspective. Requesters, who are respon-

sible for task design, seek to achieve high-quality results from crowd workers. Previous research in the field of crowdsourcing has predominantly focused on the quality of solutions provided by crowd workers (Kittur et al., 2013). It is widely acknowledged that low-quality solutions pose a substantial barrier to realizing the full potential of crowdsourcing, and unclear task design is recognized as a significant factor affecting the quality of crowd work. Therefore, ensuring clear task descriptions is imperative for achieving a high-quality task design.

Requesters are expected to provide comprehensive information in task descriptions, including required resources, steps to follow, the expected solutions and more. However, this task becomes challenging, particularly for micro-tasks that target a diverse pool of workers with varying skills, cultural backgrounds, and educational levels. Requesters with limited crowdsourcing experience often struggle to effectively describe the task requirements.

Task Clarity Assessment

Task description clarity relates to the comprehensibility and completeness of the instructions written in natural language, and also refers to the level of detail provided by requesters in order to receive a high-quality solution to the task. The study detailed in Chapter 4 aimed to investigate the effectiveness of natural language processing methods in detecting prevalent clarity issues in task descriptions, and the influence of instruction's textual properties on unclear task and associated flaws.

We envisioned technological solutions that automatically assist requesters in improving the quality of their descriptions before posting them on the platform. In contrast to existing approaches addressing ambiguous task descriptions in crowdsourcing, our solution operates independently of workers' interaction, offering a more efficient approach that saves time and money for both workers and requesters while ensuring explicit task descriptions. Moreover, this approach circumvents various challenges associated with requester-worker communication difficulties. However, the development of such tools has been limited due to the absence of effective computational methods for assessing the clarity of task descriptions.

To construct computational models for the assessment of task clarity, we drew insights from the literature focusing on the attributes of unclear and incomplete task descriptions in crowdsourcing. Initially, we pinpointed eight clarity flaws that functioned as indicators of the clarity of task instructions. We hypothesized that all these flaws negatively impacted workers' comprehension of task instructions. As there was a lack of available data for studying task clarity, we extended an existing dataset comprising 1332 real-world micro-task descriptions. These eight identified clarity flaws served as the basis for our annotation guidelines in the creation of the required dataset for model development. Ultimately, we leveraged the dataset to build two natural language processing approaches: BERT (utilizing the pre-trained

Bert-base-uncased and Bert-base-cased models) and linear SVM models with six different feature types.

Following an evaluation of the models' performance, we discovered that both approaches demonstrated effective learning capabilities for most clarity flaws, excluding difficult wording. Besides, SVMs displayed superior performance compared to the BERT models in assessing task clarity using the features we defined based on domain-specific knowledge. Furthermore, our observations indicate that the outcomes from individual feature types imply that numerous of the considered textual characteristics hold relevance to various clarity flaws. Notably, the content of task descriptions emerges as particularly significant, achieving notably higher scores than other feature types for several clarity flaws. Besides, style, and readability of descriptions are particularly important textual properties for clarity assessment. Combinations of flaw-specific properties with other features also proved advantageous for assessing clarity. However, we did not find the length of descriptions to be helpful in identifying clarity flaws.

Altogether, this study's principal contribution includes a dataset of 1332 real-world task descriptions annotated by actual workers for examining task clarity in crowdsourcing. Additionally, the study developed a feature-based and a neural approach for computationally evaluating task clarity. Lastly, it offers comprehensive empirical insights into factors that aid the computational assessment of task clarity and the extent of their impact.

Based on these insights, we proceeded to investigate whether an automated writing assistance that automatically identifies potential clarity issues in task descriptions can help requesters to improve their task description clarity. In the following, we summarize our approach to address this question.

Automated Writing Assistance

We hypothesized that an automated approach for analyzing task descriptions and predicting potential clarity flaws can help requesters to improve their task description clarity. However, to the best of our knowledge, such an approach has not been developed due to a lack of reliable computational methods for assessing task clarity. Hence, we developed an approach in Chapter 5 to study the extent of effectiveness of an automated writing assistance in helping requesters to identify clarity flaws in their task descriptions, and the extent of effectiveness of such an assistance in enhancing the task clarity for workers.

With this objective in mind, we conceived a web-based tool that empowers requesters to iteratively assess the clarity of their task instructions. Using this tool, requesters can detect the defined clarity issues in their descriptions and make necessary enhancements before posting them on the crowdsourcing platform. In the development of this tool, named "ClarifyIt," we initiated the process by adapting

the dataset and training feature-based SVM models that offer quantitative measures of clarity flaws. Subsequently, we implemented the tool. To assess the effectiveness of our tool, we conducted a comprehensive evaluation study involving both requesters and workers. The study aimed to determine the tool's usefulness for requesters and its effectiveness in resolving issues related to unclear task descriptions for workers.

The study with requesters found that participants who used ClarifyIt were able to improve the clarity of their task descriptions on all clarity dimensions. Although the initial knowledge of clarity aspects did not have a significant impact on the initial clarity scores, the iterative process helped participants enhance the clarity of their descriptions. The requesters' questionnaire responses indicate that the most appreciated feature of ClarifyIt is its general functionalities, with clarity metrics information and clarity dimensions characterization following as the next most appreciated features. The requesters perceived the tool as well-designed and effective in identifying and enhancing the clarity of task descriptions.

The study with workers indicated the best-score version of task descriptions created by requesters was clearer than the initial version in terms of all clarity dimensions. The most notable enhancements were in the specification of the desired solution, specification of task steps, overall clarity, and definition of important terms, as judged by all voters. Additionally, majority of voters observed improvements in other clarity dimensions. However, the wording and phrasing dimension received the highest percentage of negative and neutral votes, corroborating findings from the assessment study in Chapter 4, which demonstrated the difficulty of computationally assessing this particular clarity flaw.

In this stage, our regression models predicting the clarity scores of crowdsourcing task descriptions and ClarifyIt, the assistance tool to aid requesters improve their task description clarity are our major contributions to the advancements in the development of techniques for augmenting task clarity in crowdsourcing marketplaces. Overall, our evaluation study indicates that ClarifyIt effectively assists requesters in improving the clarity of their task descriptions. On average, all clarity flaws in task descriptions written using ClarifyIt demonstrated significant enhancements. This outcome suggests that the developed tool holds the potential to significantly benefit real-world crowdsourcing task descriptions.

In the following sections, we will discuss the limitations inherent in our methodologies across all stages of our work and explore avenues for future work. By addressing these aspects, we aim to provide a comprehensive perspective on our study and pave the way for future investigations and improvements in crowdsourcing marketplace.

6.2 Limitations

This section addresses the constraints associated with the approaches we employed at various stages of this thesis, which may impact our research findings. In the subsequent sections, we provide a detailed explanation for each of these limitations.

Crowdsourcing Challenges In our comprehensive study (detailed in Chapter 3) of existing challenges in crowdsourcing processes that workers face, we could not identify any problems in the data analysis related explicitly to *task operation* or *platform mediation*. While the reviews may still mention these issues, the limitations of topic modeling may cause the identified problems to take precedence and overshadow them.

Since our corpus includes discussed problems until November 2018, it is significant to emphasize that the findings of our study are inevitably merely a snapshot of the problems with crowdsourcing procedures during the period covered. To the best of our knowledge, nevertheless, task design and requesters-workers communication have not significantly altered since that time. Also, no possible solutions for fixing the stated challenges have been implemented on MTurk. Therefore, we are confident that most insights acquired from the analysis still apply to the current problems in crowdsourcing processes.

Annotated Task Descriptions A significant constraint in our research revolved around the dataset's size and quality. We required real-world task descriptions annotated for pre-defined eight clarity dimensions by crowd workers. The annotation process was a pivotal aspect of our study, demanding comprehensive coverage of crowdsourcing task description clarity dimensions, high-quality work from the annotators, and a substantial budget to create a large collection of annotated task descriptions. Regarding the comprehensive dimensions of description clarity, we integrated our domain-specific expertise with clarity dimensions outlined in the literature. This synthesis resulted in identification of eight clarity dimensions within task descriptions. The comprehensibility of these dimensions could be attributed to the varying coverage levels in the literature or our potential gaps in understanding the subject matter. Additionally, due to budget limitations, we were not able to include a large number of task descriptions in our annotation task. Besides, we were required to employ techniques to identify and eliminate low-quality submissions by annotators. The effectiveness of applied techniques directly influenced the models' performance and, consequently, the overall success of the approach in enhancing task clarity.

Automated Writing Assistance's Models In this research, one of our initial objective was to investigate how the textual characteristics of descriptions impact the

computational assessment of clarity flaws. To address this, on one hand, we initially engaged in feature engineering based on our domain-specific knowledge, and then we utilized feature-based Support Vector Machines (SVMs) to construct required models. On the other hand, the results obtained from our study in Chapter 4 indicated that, given the specific research problem and the constraints of the available annotated dataset, SVMs outperformed transformer-based algorithms. Consequently, our focus remained on the use of traditional feature-based SVMs to develop the necessary models. While these SVMs demonstrated effectiveness for our specific research objectives, it is important to note that manual feature engineering has its constraints. It may not capture the entirety of pertinent information within the data and can hinder the algorithm's capacity to autonomously learn intricate patterns and representations from the training data. In contrast, state-of-the-art transformer-based algorithms offer a different set of advantages. These encompass automatic feature learning and the capability for enhanced generalization, particularly in scenarios where SVMs may be less versatile when applied within a broader context.

Automated Writing Assistance's Evaluation In the context of the user study with requesters, the ideal scenario would have been to directly engage with real-world, active requesters and enlist their participation in our research. Unfortunately, establishing connections with a significant number of requesters and coordinating the necessary arrangements for their involvement proved to be an impractical endeavor. However, this could have a promising impact on the insights gained from their feedback on the helpfulness of our tool.

6.3 Future Work

Our work opens up several promising avenues for future research. In the following, we will provide a brief discussion on each topic.

Challenges from Requesters' Perspective Our study conducted in Chapter 3 offered an extensive analysis of the most frequently encountered challenges faced by crowd workers in their daily work within crowdsourcing processes. These findings may be valuable for requesters, crowdsourcing platform providers, and researchers interested in understanding the challenges elaborated on in both scholarly literature and real-world worker experiences. Although a comparable in-depth examination of the problems from the requesters' perspective would provide valuable insights, to the best of our knowledge, there was a lack of literature and data available for this purpose at the time of our work.

Dataset One way to enhance the performance of computational models is by expanding the dataset used for training the models to include a more diverse range of micro-task descriptions that have been annotated for clarity flaws. By incorporating a larger and more varied dataset, it may be possible to train models that are more accurate and precise in their predictions.

Difficult Wording in Descriptions In terms of task descriptions clarity flaws, the challenge of identifying the clarity of wording in task descriptions appears to be particularly difficult. To improve the prediction of this specific dimension, it may be necessary to develop more refined models that can effectively capture the subtleties and ambiguities often present in task descriptions. The following point has the potential to contribute to enhancements in predicting difficult wording flaw in task descriptions.

Computational Models Regarding models, it would be valuable to evaluate the effectiveness of the latest transformer-based models, such as DeBERTA (He et al., 2020, 2021), in computational assessment, to determine if they can further enhance the detection of clarity flaws in crowdsourcing task descriptions.

Moreover, it would be beneficial to explore Large Language Models (LLMs) to build the computational models capable of identifying clarity issues in task descriptions. LLMs are artificial intelligence-powered systems that are adept at producing text which closely resembles human writing. These models learn from a substantial corpus of text data, and this acquired knowledge empowers them to craft responses or generate new content sparked by the prompts they encounter (Radford et al., 2018). A prominent example of an LLM is OpenAI's Generative Pre-trained Transformer (GPT) and ChatGPT is a specialized variant of the GPT model that is fine-tuned for generating conversational responses. The integration of the ChatGPT API into our tool can enhance its effectiveness to assist users in identifying clarity flaws within their task descriptions. Regarding the challenge of handling difficult wording dimension of descriptions, this approach could make the most of ChatGPT's ability to comprehend language, locating vague or confused expressions, proposing more accurate phrasing, and offering advice on augmenting the structure of the entire content. Furthermore, capitalizing on ChatGPT's abilities in understanding and generating responses aware of their context, the tool could deliver customized enhancements in a specific task description.

Tool's Improvements In terms of the tool's helpfulness, some requesters expressed their desire for real-time feedback on clarity while they are typing a task description. This feature would enhance the efficiency and effectiveness of the process. Additionally, requesters expressed interest in receiving suggestions for improving the clarity of their descriptions, similar to the functionality provided by tools

like Grammarly. Another feature that requesters found valuable is an undo button, which would allow them to revert to previous versions of a task description in case changes result in lower clarity scores. Overall, we believe that integrating ClarifyIt as a plug-in tool on crowdsourcing platforms could bring significant benefits to requesters by providing them with automated assistance in writing clear and concise task descriptions. It is also worth noting that our computational approach can still be complemented by workflows or interventions that incorporate the opinions and feedback of workers if deemed necessary.

Effect of Task Description Clarity on Quality of Submissions Exploring the relationship between improvements in key dimensions of effective task design and the quality of final results would be beneficial. This involves the need for a sophisticated user study design where pairs of tasks are presented. These tasks should be well-designed, considering factors like fair time and effort estimation as well as fair compensation. In each pair of tasks, the task instructions should only vary to carefully assess how task clarity impacts result quality. Additionally, the expected results should be straightforward and easy to understand to enable a fair comparison between the quality of submissions and the anticipated outcomes.

Other Domains When considering the application of a similar method to the other domains or fields of study, it is important to assess how the approach can be adapted and utilized effectively in those diverse contexts. However, we believe that similar approaches to providing automated support for textual content creators could be explored in other domains. Tools that assist creators in identifying and improving clarity flaws in their texts based on operationalizable clarity specifications could be developed.

As there are multiple factors that contribute to effective task design, including factors like fair time and payment estimation, a well-designed feedback system, and clarity of task instructions, it is beyond the scope of our study to directly evaluate the impact of ClarifyIt on the quality of workers' final results. However, we can draw insights from previous research, such as the study conducted by Wu and Quinn (2017), to provide a broader perspective on whether our work can be considered a practical approach towards improving the quality of workers' final results, which is a high-level goal. Wu and Quinn (2017) highlighted the significance of task instruction clarity in influencing workers' behavior, emphasizing that requesters should have a good understanding of task requirements and the principles of task description design. The findings from our evaluation study align with this, demonstrating that ClarifyIt effectively aids requesters in comprehending and addressing clarity flaws in their task instructions. This is crucial in tackling the issue of low-quality submissions by workers and improving crowdsourcing processes overall.

Appendix A

Example Tables

In the upcoming sections, we furnish additional examples representing instances gathered during the thesis process. Section A.1 offers sample tables pertaining to the study of crowdsourcing challenges carried out in Chapter 3, while Section A.2 showcases a sample table containing real-world task descriptions that have been enriched with annotations pertaining to eight distinct clarity dimensions. These annotations are essential for the assessment study conducted in Chapter 4. Section A.3 presents sample tables associated with the evaluation of our tool's effectiveness conducted in Chapter 5.

A.1 Crowdsourcing Challenges

In this section, we provide additional example instances associated with the study of crowdsourcing challenges conducted in Chapter 3. Section A.1.1 offers a list of the papers that were excluded during our literature review, while Section A.1.2 presents additional review examples on Turkopticon platform utilized in our empirical data analysis.

A.1.1 Literature Review

Here, we provide a list the 69 papers that were found during our search process in the literature review but excluded due to their misalignment with the study's scope. We provide the an overview on those papers in two tables. Table A.1 classifies the domains and subjects we explored during the paper review, along with the respective counts of papers within each category of emphasis.

As indicated in Table A.1, the papers that were excluded predominantly fell within the fields of *information systems and human computation*, *psychology*, *business and organizational management*, as well as law and sociology, all of which were categorized under the *other* classification.

Additionally, Table A.2(a) presents a list of 31 identified papers within the domain of information systems and human computation, concentrating on topics

Area	Focus	# Papers
Information Systems and Human Computation	Models/Methods/Benefits/Risks	31
	Quality Management	19
Psychology	Behavioural Job Attitudes	6
Business and Organization Management	Workforce Management Models	10
Other	Law / Sociology	3

TABLE A.1: This table presents the 69 papers categorized into various research domains. These papers underwent a literature review process, but they were omitted from our study as they did not align with the study's scope due to their specific focus.

related to crowdsourcing methods, models of crowdsourcing within this field, and the associated risks and benefits. Furthermore, the table includes 19 papers that center on quality management strategies for enhancing crowdsourcing models in the context of information systems.

Furthermore, Table A.2(b) provides a compilation of 6 identified papers in the field of psychology, with a focus on subjects concerning crowd workers and their rights within the context of crowdsourcing processes. Researchers also explore the attitudes of workers in crowdsourcing in comparison to traditional work models.

Table A.2(c) also offers a summary of 10 located papers within the realm of business and organizational management. These papers emphasize the management of the workforce in the context of crowdsourcing and the necessary adaptations in management approaches to achieve success in the business domain.

Lastly, table A.2(d) offers a summary of 3 located papers in various fields like law and sociology and application of crowdsourcing work model in the areas.

Focus	# Papers
(a) Information Systems and Human Computation	
Models/ Methods/ Benefits/ Risks	(Gelderman, 2002, Snow et al., 2008, Brabham, 2008a, Kazai, 2010, Huang et al., 2010, Euchner, 2010, Chanal and Caron-Fasan, 2010, Zheng et al., 2011, Doan et al., 2011, Geiger et al., 2011, Harris, 2011b, Dai et al., 2011, Schenk and Guittard, 2011, Harris, 2011a, Varshney, 2012, Brabham, 2012, Zhang and van der Schaar, 2012, Anya et al., 2013, Kittur et al., 2013, Saxton et al., 2013, Simula, 2013, Hosseini et al., 2014, Sherief et al., 2014, Zhao and Zhu, 2014, Simperl, 2015, AyferBozat and Erenel, 2016, Wu and Quinn, 2017, Feng et al., 2017, Yang et al., 2018, Liu et al., 2018, Manam and Quinn, 2018).
Quality Management	(Lakhani and Wolf, 2003, Wise et al., 2006, Brabham, 2008b, Johnston et al., 2009, Ipeirotis, 2010, Wais et al., 2010, Wang et al., 2011, Kazai et al., 2011, Koch et al., 2011, Gutheim and Hartmann, 2012, Gawade et al., 2012, Huang and Fu, 2013, O'Neill and Martin, 2013, Allahbakhsh et al., 2013, Della Mea et al., 2013, Straub et al., 2014, Varshney et al., 2014, Lasecki et al., 2015, Gray et al., 2016, Wijermans et al., 2016)
(b) Psychology	
Behavioural Job Attitudes	(Betsch et al., 1998, Tetrick et al., 2000, Cordery et al., 2010, Judge and Kammeyer-Mueller, 2012, Templer, 2012, Mason and Suri, 2012)
(c) Business and Organization Management	
Workforce Management Models	(Campbell, 1988, Jang et al., 2008, Thompson and Phua, 2012, Djelassi and Decoopman, 2013, Kaganer et al., 2013, Bourne and Forman, 2014, Buettner, 2015, Verschoore et al., 2015, Ryan and Wessel, 2015, Osnowitz and Henson, 2016)
(d) Other	
Law / Sociology	(Kalleberg, 2009, Aloisi, 2015, Zhang et al.)

TABLE A.2: This table provides the list of 69 papers classified based on their research domains.

A.1.2 Data Analysis

for our empirical data analysis conducted in Chapter 3, we selected Turkopticon due to its explicit focus on advocating for workers' rights in their interactions with requesters. Turkopticon serves as a reputation system where workers provide reviews, detailing their experiences related to tasks, payments, and rejections encountered during their daily work. Within these reviews, workers assign ratings to requesters and express whether they recommend or do not recommend a particular requester to their fellow workers. This information allows other workers to make informed decisions by reading these reviews before accepting tasks from requesters.

Table A.3 provides 20 sample out of total 8610 reviews tagged as "not recommended," where workers shared their negative experiences with requesters. We utilized these reviews to extract crowdsourcing challenges discussed by Turkopticon users.

TABLE A.3: 20 example reviews labeled as "not recommended" from Turkopticon that were employed for the empirical data analysis. Note: Tasks on MTurk platform are called HITs which stands for Human Intelligence Tasks.

#	Example Reviews
1	"This survey is very personal and can be distressing to some. It has a lot of bubbles and then quite a bit of reading. I feel it should pay quite a bit more for the information they are looking for. He does state this on the consent page, but sometimes we skim those."
2	"Very monotonous. Listen to excerpts of music in various languages and then bubble hell over and over and over. Supposedly a $2.75 bonus but still not worth it even with the bonus."
3	"Asks for too much personally identifying info. There was an unpaid screener. This hit asked for entirely too much personally identifying information, I returned it."
4	"This requester's HITS are hit and and miss. This one was a real miss. 14 pages long with scroll down buttons to boot. 24 hour AA"
5	"Answered some questions about past traumas. Took me 20 mins. I'm sure someone else could do it faster. Some light writing, then bubble up your seatbelts, haha."

Continued on next page

TABLE A.3: 20 example reviews labeled as "not recommended" from Turkopticon that were employed for the empirical data analysis. Note: Tasks on MTurk platform are called HITs which stands for Human Intelligence Tasks.

#	Example Reviews
6	"This requester has been getting worse and worse. I only did this HIT because I've had a really slow week. Used to be one of my favorite requesters I would always look for, now I only do them when I'm desperate."
7	"Its long and the bubbles are never ending. Just about fair for a $4 hit but tiresome, lots of repetitive questions. Chance for a second follow up hit worth $5"
8	"This requester did not pay me, nor did they answer my request for payment, A NON PAYER! Terrible"
9	"The HIT says it takes 15-25 minutes so I decided to risk it, since I'm usually faster than the estimates. It took me the whole 25 and involved memorizing graphs. Way too brain intensive for something underpaid."
10	"tos requires downloading an .exe file to do the experiment.,not recommended TOS Violation. [It] requires downloading an .exe file to do the experiment. plus it says it will take 20-60 minutes to do the experiment .. for $1."
11	""In this study, you will be asked questions related to your racial/ethnic identity, self-concept, experiences of insults, and sexual health. The full study will take about 60 minutes, and you will receive $1.00 in Amazon credit. To access survey, you must enter your personal, unique survey password.",Doesn't tell you where to get said password. I'm guessing there was another pre-screen HIT, but unqualified people (like me) can accept this follow up HIT.,Kind of glad honestly. Based on her other reviews ([provided link]), it would probably end up taking the actual 60 mins and then some for a lousy buck."
12	"Stopped after about 13 mins and several timed pages. Even timed pages of bubbles where you sit waiting for the timer, ugh."
13	"A series of 7 surveys over 20 business days for a total of $21.00. That comes out to $1.05 a day when all done. Pretty low. It seems really tedious as well."
14	"Bubbles, not terribly long though. Pay is meh."

Continued on next page

TABLE A.3: 20 example reviews labeled as "not recommended" from Turkopticon that were employed for the empirical data analysis. Note: Tasks on MTurk platform are called HITs which stands for Human Intelligence Tasks.

#	Example Reviews
15	“Rejected immediately with this response:,"Please note that workers who completed my first survey in September 2018 can participate in this survey and get credit. Since you are not qualified for this survey, I have to reject. I am very sorry about that.",Not only does the explanation not make sense, but the HIT instructions clearly state:,"1. We are conducting an academic survey about mobile loafing. We need to understand your opinion about mobile Internet use at work.”
16	“10 minutes for $0.50 = $3.00/hr”
17	“Requires Inquisit to complete.,([provided link]),They keep trying to raise the pay to get people to take the survey, but according to reviews it’s still underpaid at $3.00.”
18	“New requester, did not provide code, was not academic survey,ETA: was paid fine. Google docs survey.”
19	“needs Adult Content qual, doesn’t have it, not recommended. pay should be higher for something this long”
20	“an unpaid screen of several pages that was basically a full survey of information...complete scammer...blocked him”

A.2 Task Clarity Assessment

In order to facilitate the task clarity assessment study conducted in Chapter 4, we undertook a four-step process to create and validate a dataset. Table A.4 provides 20 task description instances derived from the dataset creation process.

TABLE A.4: 20 sample task descriptions from the dataset, each annotated by crowd workers based on eight defined clarity dimensions. The ratings of one to five for each dimension indicate the level of clarity in the respective aspect. A rating of "5" signifies a **very clear** task description, "4" denotes **clear**, "3" signifies **partial (un)clarity**, "2" indicates **unclear**, and "1" reflects **very unclear**. This Dataset were used for building the models in this thesis.

#	Task Description	Overall clear	Easy wording	Terms def.	Solution spec.	Format spec.	Steps spec.	Resources spec.	Criteria spec.
1	**Title:** Just click on the link and vote for us **Body:** If you have a valid Facebook account, just click on the this link: 'the given link' and click on the Vote Now button. Confirm by logging into to Facebook and you're done.	5	5	5	4	5	5	5	4
2	**Title:** What is the Best House for Trick or Treating in your neighborhood and why is it the best? **Body:** We are building a neighborhood gossip and information site and would like to know about what is going on around you.	5	5	4	5	5	4	5	5
3	**Title:** Vote for My Sweater Design **Body:** Please go to this link: 'a given link' Sign in with your Facebook or Twitter account. Make sure the page says "Robin" at the top and has a sweater with dinosaurs on it. Click the mitten inside of the red circle to vote for the sweater.	5	1	1	4	2	2	2	4

Continued on next page

TABLE A.4: 20 sample task descriptions from the dataset, each annotated by crowd workers based on eight defined clarity dimensions. The ratings of one to five for each dimension indicate the level of clarity in the respective aspect. A rating of "5" signifies a **very clear** task description, "4" denotes **clear**, "3" signifies **partial (un)clarity**, "2" indicates **unclear**, and "1" reflects **very unclear**. This Dataset were used for building the models in this thesis.

#	Task Description	Overall clear	Easy wording	Terms def.	Solution spec.	Format spec.	Steps spec.	Resources spec.	Criteria spec.
4	**Title:** Free Easy Sign Up! Only enter Name, Email, DOB and Gender. (18+ only, Adult Content) **Body:** Free Easy Sign Up! Only enter Name, Email, DOB and Gender...	5	1	1	4	1	2	3	2
5	**Title:** Find "link bracelet" on A ma zon dot com **Body:** Just a few steps to complete this task	4	4	5	5	5	4	4	5
6	**Title:** Data entry from business card images **Body:** Type business card information	4	4	5	5	5	5	4	5
7	**Title:** Please review: Search :Keywords on Google.com(US) and report results **Body:** Review a hit with the following description: Search a keyword in Google and report the results Keywords	4	5	5	4	4	4	2	2
8	**Title:** Give us a vote - one click, super easy **Body:** Vote for us - just one click	4	3	3	2	2	3	3	3
9	**Title:** Search through Satellite Imagery to find various items **Body:** Search through Satellite Imagery to find various items	3	1	5	5	4	5	3	5
10	**Title:** Contact information for this business person								

Continued on next page

TABLE A.4: 20 sample task descriptions from the dataset, each annotated by crowd workers based on eight defined clarity dimensions. The ratings of one to five for each dimension indicate the level of clarity in the respective aspect. A rating of "5" signifies a **very clear** task description, "4" denotes **clear**, "3" signifies **partial (un)clarity**, "2" indicates **unclear**, and "1" reflects **very unclear**. This Dataset were used for building the models in this thesis.

#	Task Description	Overall clear	Easy wording	Terms def.	Solution spec.	Format spec.	Steps spec.	Resources spec.	Criteria spec.
	Body: For the person below, find and enter their email address. (Include the full email address.)	3	5	4	5	5	5	4	5
11	**Title:** Edit an Expedited Transcript (Denver Thomas Roger #####) (avg rwrd+bns: $2.58) **Body:** Edit a Difficult Audio Transcript: 'Denver Thomas Roger #####'	3	2	3	5	4	5	3	4
12	**Title:** World Vision: Approve or Reject Greeting Videos **Body:** Approve or reject images and videos for posting on website	3	2	4	5	5	2	2	3
13	**Title:** Take a chance on this survey! **Body:** This survey will ask you how you think about how quizzes	2	3	2	4	4	3	3	4
14	**Title:** Find the Email Addresses of News Reporters **Body:** Given the website article link, find the official email address of these news reporters	2	2	3	3	5	4	3	3
15	**Title:** Find redundancies in arguments about Gun Control **Body:** Given two collections of arguments about the political topic Gun Control, your task will be to identify and mark redundancies between arguments.	2	1	1	2	4	2	1	2

Continued on next page

TABLE A.4: 20 sample task descriptions from the dataset, each annotated by crowd workers based on eight defined clarity dimensions. The ratings of one to five for each dimension indicate the level of clarity in the respective aspect. A rating of "5" signifies a **very clear** task description, "4" denotes **clear**, "3" signifies **partial (un)clarity**, "2" indicates **unclear**, and "1" reflects **very unclear**. This Dataset were used for building the models in this thesis.

#	Task Description	Overall clear	Easy wording	Terms def.	Solution spec.	Format spec.	Steps spec.	Resources spec.	Criteria spec.
16	**Title:** Extract purchased items from Walmart shopping receipt **Body:** Transcribe UPCs and amounts from a grocery receipt	2	2	2	2	3	2	3	2
17	**Title:** Tell us the address after clicking on the links name **Body:** Click link and tell the address	1	2	2	2	2	3	2	3
18	**Title:** Help us test our website's user interface **Body:** Perform some simple tasks to help test how easy our site is to use	1	1	1	4	2	4	4	3
19	**Title:** Watch then upvote youtube video **Body:** Watch then upvote youtube video	1	2	3	2	2	3	3	3
20	**Title:** 5-10min task Economics and America **Body:** Short Survey for $1.50	1	1	1	2	1	3	3	2

A.3 Automated Writing Assistance

In the user study involving requesters, participants were instructed to utilize our tool. They were tasked with creating a task description, complete with a title and instructional content, based on a provided scenario. Subsequently, they were asked to assess the clarity of this description using our tool and make improvements guided by the information provided by the tool. Participants were encouraged to repeat these steps in a series of iterations until they achieved a level of description clarity that met their satisfaction, as indicated by the clarity scores displayed on the tool.

In the upcoming sections, we will present detailed tables featuring examples obtained from the assessment study with requesters as conducted in Chapter 5. Section A.3.1 provides examples for feedback from participants, while Section A.3.2 presents example instances of the task description pairs created by participants in the evaluation study.

A.3.1 Feedback on ClarifyIt

In the final phase of the user study involving requesters, participants from two main groups, nameky researchers and crowd workers on Prolific platform, had the opportunity to provide feedback about their experience using our tool to enhance the clarity of their task descriptions. This feedback was collected through an open-text field. Table A.5 displays 20 examples of the feedback written by participants who took part in the requester evaluation study.

TABLE A.5: 20 sample comments obtained during the user study with requesters for the assessment of the tool's utility. These comments were selected from both requester groups, which consisted of researchers and Prolific workers who participated in the user study.

#	Group	Comment
1	Researcher	There is no way to know how the AI assess the writing based on the metrics (I think my writing is improved but did not see changes in the AI evaluation). There is no direct feedback (e.g., I personally love the grammarly way to engage writers). Score does not help since my goal is to get a better task, not to maximize the score. If there is no feedback about how to improve, then it is not very useful to me, as I can just search how other people improve their task and use these guidelines to improve my descriptions offline. Also, the AI took too long to evaluate the writing, and I have no patience to wait for that long (grammarly gives almost instant feedback).

Continued on next page

TABLE A.5: 20 sample comments obtained during the user study with requesters for the assessment of the tool's utility. These comments were selected from both requester groups, which consisted of researchers and Prolific workers who participated in the user study.

#	Group	Comment
2	Researcher	I think it would be helpful extension if you could show which words and/or sentences influences the score for each metric the most.
3	Researcher	The idea and implementation of the tool was amazing and would like to definitely try it in the future. I might more helpful to highlight the problematic part on bad examples. Especially highlighting problematic parts of the main task would be super helpful or a description of why we get specific score. Because for me it was difficult to understand why after improving the text based on the mentioned criteria I still get low scores. Thank you for this interesting tool.
4	Researcher	It was not clear to me that the bad and good example buttons could be used to show bad or good examples. Instead, I felt like it would be used to rate the description of the clarity metric. I think this was because I first clicked the arrow to expand the clarity metric description and only then saw the circle buttons. I also didn't see an explanation of these buttons. So I actually didn't use the example buttons and feel like they could have helped me after completing this survey.
5	Researcher	I'm afraid, after a few tries of changing the wording of my task description (e.g., trying shorter ones, trying longer ones, adding sentences addressing each clarity metric, etc.), I could not figure out a way of systematically increasing my score. Perhaps the words "health" or "web page" or "website" were in general judged too "complicated" by the tool? There was no indication or hint as to where exactly the model thought that something was unclear, which left me randomly guessing. This is of course frustrating. Since I do think I know how to describe a task, and have improved task descriptions in sometimes tedious baby steps and discussions, I still felt "left alone" with these abstract numbers. Perhaps the models could highlight phrases which it most thinks are bad?

Continued on next page

TABLE A.5: 20 sample comments obtained during the user study with requesters for the assessment of the tool's utility. These comments were selected from both requester groups, which consisted of researchers and Prolific workers who participated in the user study.

#	Group	Comment
6	Researcher	All metrics showed 100% in the first place, so I was somewhat confused on whether that was the actual truly assessments by the tool or I just messed something up.
7	Researcher	The metrics although helped me notice what I was missing, I believe the scores needed more explanation. I was surprised I got low clarity scores after adding clarifying details on the task. Will be helpful to know what is missing maybe.
8	Researcher	Very interesting tool! My main concern was that no matter what changes I did, the feedback was almost always the same. The overall clarity was around 34% regardless of changes to the structure of the task. I rewrote it from a narrative to a more structured description with clear labels. However, not a lot changed in terms of scores (even individual metrics were in the same range) and thus I could not identify my mistakes. I was still left guessing on what to do to improve the score. May be it was specific to the task I envisioned and described. This is a very challenging task indeed and I appreciate the effort to tackle it! Good luck.
9	Researcher	I would suggest to describe the metrics as categories: bad, average, excellent; I did not how to interpret the scores and when to stop improving.
10	Researcher	I had very little change in the metrics even with huge changes in the text, and it was very hard to understand how my changes were impacting the metrics. After some attempts, it felt kinda random. I know it's difficult, and maybe out of scope, but it would be great to indicate the text itself (like this word or sentence is too difficult, or this part of the text contributes to this metric in this way). Nevertheless, overall it's an interesting work!!
11	Researcher	I used the tool a bit like a checklist. However, the metrics did not change too much after I have made quite a lot of changes. This was a bit frustrating.

Continued on next page

TABLE A.5: 20 sample comments obtained during the user study with requesters for the assessment of the tool's utility. These comments were selected from both requester groups, which consisted of researchers and Prolific workers who participated in the user study.

#	Group	Comment
12	Researcher	It would be useful to highlight problem areas and give more specific feedback and examples (which parts of the description are not phrased simply? which terms are missing its definitions? how to specify steps to perform tasks?). Right now it reads more like a high-level, general checklist or reminder what to focus on when proof-reading your own description.
13	Researcher	There is little to no transparency on why a certain change in the title or description leads to a change in a certain metric. There is no way to compare multiple versions of the title and description, this might be useful to spot differences in text and how that affects the score.
14	Prolific Worker	Initially after the first couple of iterations the text was definitely better however after this it was very difficult to improve and on a couple of occasions the metrics went down. This was very frustrating as the modifications were better for some metrics but worse for others.
15	Prolific Worker	The tool was extremely helpful. I lacked imagination in creating the task specifics however the metrics were genuinely helpful in clarifying what workers would need to know.
16	Prolific Worker	This was very interesting to do, especially as I have not crowdsourced before, but I now understand more of what I wold have to do and this tool would be useful as part of the process.
17	Prolific Worker	I wasn't sure of the difference between the 'clarity' measurement and the 'AI confidence' measurement.
18	Prolific Worker	It helps me get the wording right so it is reaching the correct audience
19	Prolific Worker	It is useful for picking up typos and if I have used awkward wording. It feels like it's expecting a giant wall of text that people who don't take surveys won't read or could be better short for simpler surveys.

Continued on next page

TABLE A.5: 20 sample comments obtained during the user study with requesters for the assessment of the tool's utility. These comments were selected from both requester groups, which consisted of researchers and Prolific workers who participated in the user study.

#	Group	Comment
20	Prolific Worker	I found the idea of the tool interesting, but I don't think there was enough specific feedback that helped me work out how to improve. So it was a bit of trial and error to improve the query. Sometimes things that I did that I thought were good seemed to make the scores go down, so can't really explain that very well.

A.3.2 Task Descriptions

Table A.6 shows 10 instances of task description pairs, comprising the original version and the highest-scoring version, along with the respective iteration number in which the highest-scoring version was improved. These pairs are derived from a pool of 100 pairs created during the requester user study and were subsequently employed in the worker user study to evaluate the tool's effectiveness from the workers' perspective.

TABLE A.6: 10 instances of task description pairs, comprising the original version and the highest-scoring version, along with the respective iteration number in which the highest-scoring version was improved using our tool.

#	Iters	Initial Task Description	Best Task Description
1	10	**Title:** Crowdworking questionnaire **Body:** Please answer this questionnaire on this latest tool used for Crowd working.	**Title:** Tools used when crowd working **Body:** This is about the latest tools. You may see these used for Crowd working. Please answer to the best of you ability.
2	11	**Title:** Text writing tasl **Body:** Your task is to write content on an arbitrary topic.	**Title:** Text writing task **Body:** Your task is to write clear and concise content on a chosen topic. You will have fifteen minutes to write 2000 words on the chosen topic. You will be given a textbox to write this in. Spelling, punctuation and grammar is important. Your payment for completing the task will be £1.
3	2	**Title:** Assess a piece of text **Body:** Read and feedback on a piece of text.	**Title:** Assess a piece of text **Body:** Read and feedback on a piece of text. Your thoughts are valuable to us.
4	7	**Title:** Feedback required on content **Body:** Hello everyone, I need you to give feedback on this content please	**Title:** Feedback of Crowd workers **Body:** The purpose of this survey is to obtain feedback from crowd workers regarding the text or content involved.

Continued on next page

TABLE A.6: 10 instances of task description pairs, comprising the original version and the highest-scoring version, along with the respective iteration number in which the highest-scoring version was improved using our tool.

#	Iters	Initial Task Description	Best Task Description
5	4	**Title:** Sponsor a Kite Festival **Body:** A new kite festival is planned for 2023 and the aim is to have as many international guests as possible from all or even every country in the world. To do this funding is needed to support the travel costs and accommodation costs of the event.	**Title:** Sponsor a Kite Festival **Body:** A new kite festival is planned for 2023 and the aim is to have as many international guests as possible from all or even every country in the world. To do this funding is needed to support the travel costs and accommodation costs of the event. As people will be travelling long distances in some cases, additional support is needed to provide accommodation to these guests before and after the event. The target figure is £500,000 which covers the abopve and provides support for 250 kite-fliers from around the world. The sum also provides the costs for hiring of the site, all infrastructure requirements and additional costs. Support does not have to be in opure money, support in kind is also welcome - such as PA or volunteering.
6	2	**Title:** Write text on holiday destinations **Body:** You are required to write a text on different holiday destinations. The text should describe the location of the holiday and the different sights that can be seen and the cost.	**Title:** Write text on holiday destinations **Body:** You are required to write a text on different holiday destinations. The text should describe the location of the holiday and the different sights that can be seen and the cost. It should be as simple as possible

Continued on next page

TABLE A.6: 10 instances of task description pairs, comprising the original version and the highest-scoring version, along with the respective iteration number in which the highest-scoring version was improved using our tool.

#	Iters	Initial Task Description	Best Task Description
7	8	**Title:** Call for writers **Body:** Writers needed to complete a series of texts on diverse topics. Please send a copy of your recent work and apply for this role.	**Title:** Call for writers/crowd workers for a piece of texts **Body:** We are looking for Writers and crowd workers to complete a piece of texts on diverse and arbitrary topics that will be discussed after the applications have been received. Please apply for this role and send us a copy of your recent work for evaluation. You can find the application form below and also the guidelines in order to fill it out appropriately.
8	2	**Title:** Texts are wanted for a variety of topics. **Body:** Looking for experienced writers to work on a new piece of text on a variety of topics. Must be used to writing on a variety of subjects.	**Title:** Texts are wanted for a variety of topics. **Body:** Looking for experienced crowd workers with experience in writing to work on a new piece of text on a variety of topics. Must be used to writing on a variety of subjects.
9	2	**Title:** Write for $$$$$$ **Body:** Writers required to write a 500 word text on an arbitrary topic. - Research the topic at hand - Write instinctively and intuitively - Bonus for clarity	**Title:** Write for $$$$$$ **Body:** Writers needed to write a 500 word text on an arbitrary topic- a topic of your choice. If you have a flair for writing, this can be the job for you. - Research the topic at hand - Write instinctively and intuitively - Bonus for clarity - Proof read work Thank you.

Continued on next page

TABLE A.6: 10 instances of task description pairs, comprising the original version and the highest-scoring version, along with the respective iteration number in which the highest-scoring version was improved using our tool.

#	Iters	Initial Task Description	Best Task Description
10	2	**Title:** Write short plots for a role playing / action videogame **Body:** In this task you will be writing a short paragraph that contains a plot for a proposed videogame. These paragraphs should be at least 4-5 sentences and give the premise for a videogame.	**Title:** Write short plots for a role playing / action videogame **Body:** In this task you will be writing a short paragraph that contains a plot for a proposed videogame. These paragraphs should be at least 4-5 sentences and give the premise for a videogame. These plots should contain things like setting(where and when the game is being taken place), characters (protagonists/antagonists, side characters, etc), and a conflict(what is trying to be resolved in the story).

Bibliography

Joan Acker. Hierarchies, jobs, bodies: A theory of gendered organizations. *Gender and Society*, 4(2):139–158, 1990. doi:10.1177/089124390004002002. URL `https://doi.org/10.1177/089124390004002002`.

Eric Afful-Dadzie, Stephen Nabareseh, OZ Komínková, and Petr Klímek. Enterprise competitive analysis and consumer sentiments on social media. In *Proceedings of 3rd International Conference on Data Management Technologies and Applications*, pages 22–32, 2014.

Harini Alagarai Sampath, Rajeev Rajeshuni, and Bipin Indurkhya. Cognitively inspired task design to improve user performance on crowdsourcing platforms. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3665–3674, 2014.

Mohammad Allahbakhsh, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, 17(2):76–81, 2013.

Antonio Aloisi. Commoditized workers: Case study research on labor law issues arising from a set of on-demand/gig economy platforms. *Comp. Lab. L. & Pol'y J.*, 37:653, 2015.

Omar Alonso and Ricardo Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *European Conference on Information Retrieval*, pages 153–164. Springer, 2011.

Obinna Anya, Melissa Cefkin, Steve Dill, Robert Moore, Susan Stucky, and Osarieme Omokaro. Making crowdwork work: Issues in crowdsourcing for organizations. In *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.

Rajkumar Arun, Venkatasubramaniyan Suresh, CE Veni Madhavan, and MN Narasimha Murthy. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 391–402. Springer, 2010.

Zeynep AyferBozat and Fahri Erenel. Open innovation in human resources and the use of crowdsourcing. *Modern Management Science & Engineering*, 4:43, 2016.

Benjamin B Bederson and Alexander J Quinn. Web workers unite! addressing challenges of online laborers. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 97–106. ACM, 2011.

Janine Berg, Marianne Furrer, Ellie Harmon, Uma Rani, and Max Six Silberman. Digital labour platforms and the future of work. *Towards Decent Work in the Online World. Rapport de l'OIT*, 2018.

Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 313–322, 2010.

Tilmann Betsch, Klaus Fiedler, and Julia Brinkmann. Behavioral routines in decision making: The effects of novelty in task presentation and time pressure on routine maintenance and deviation. *European Journal of Social Psychology*, 28 (6):861–878, 1998.

Dhruv A Bhatt, Kristin E McNeil, and Nitaben A Patel. Time-based sentiment analysis for product and service features, November 3 2015. US Patent 9,177,554.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Mark Boons, Daan Stam, and Harry G. Barkema. Feelings of pride and respect as drivers of ongoing member activity on crowdsourcing platforms. *Journal of Management Studies*, 52(6):717–741, 2015. doi:https://doi.org/10.1111/joms.12140.

Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40, 2009.

Kristina A Bourne and Pamela J Forman. Living in a culture of overwork: An ethnographic study of flexibility. *Journal of Management Inquiry*, 23(1):68–79, 2014.

Daren C Brabham. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence*, 14(1):75–90, 2008a.

Daren C Brabham. Moving the crowd at istockphoto: The composition of the crowd and motivations for participation in a crowdsourcing application. *First monday*, 13(6), 2008b.

Daren C Brabham. The myth of amateur crowds: A critical discourse analysis of crowdsourcing coverage. *Information, Communication & Society*, 15(3):394–410, 2012.

Jonathan Bragg, Daniel S Weld, et al. Sprout: Crowd-powered task design for crowdsourcing. In *The 31st Annual ACM Symposium on User Interface Software and Technology*, pages 165–176. ACM, 2018.

Alice M Brawley and Cynthia LS Pury. Work experiences on mturk: Job satisfaction, turnover, and information sharing. *Computers in Human Behavior*, 54: 531–546, 2016.

Ricardo Buettner. A systematic literature review of crowdsourcing research from a human resource management perspective. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, pages 4609–4618. IEEE, 2015.

Donald J Campbell. Task complexity: A review and analysis. *Academy of management review*, 13(1):40–52, 1988.

Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. A density-based method for adaptive LDA model selection. *Neurocomputing*, 72(7-9): 1775–1781, 2009.

Jeanne Sternlicht Chall and Edgar Dale. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books, 1995.

Valérie Chanal and Marie-Laurence Caron-Fasan. The difficulties involved in developing business models open to innovation communities: The case of a crowdsourcing platform. *M@ n@ gement*, 13(4):318–340, 2010.

Jesse Chandler, Gabriele Paolacci, and Pam Mueller. Risks and rewards of crowdsourcing marketplaces. In *Handbook of human computation*, pages 377–392. Springer, 2013.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2334–2346, 2017.

Jenny J Chen, Natala J Menezes, Adam D Bradley, and T North. Opportunities for crowdsourcing research on amazon mechanical turk. *Interfaces*, 5(3):1, 2011.

Kevyn Collins-Thompson and James P Callan. A language modeling approach to predicting reading difficulty. In *Proceedings of the human language technology conference of the North American chapter of the association for computational linguistics: HLT-NAACL 2004*, pages 193–200, 2004.

John L Cordery, David Morrison, Brett M Wright, and Toby D Wall. The impact of autonomy and task uncertainty on team performance: A longitudinal field study. *Journal of organizational behavior*, 31(2-3):240–258, 2010.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

Peng Dai, Daniel S Weld, et al. Decision-theoretic control of crowd-sourced workflows, December 22 2011. US Patent App. 13/049,769.

Vincenzo Della Mea, Eddy Maddalena, and Stefano Mizzaro. Crowdsourcing to mobile users: A study of the role of platforms and tasks. In *DBCrowd*, pages 14–19. Citeseer, 2013.

Xuefei Nancy Deng and Kshiti D Joshi. Why individuals participate in micro-task crowdsourcing work environment: Revealing crowdworkers' perceptions. *Journal of the Association for Information Systems*, 17(10):648, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423. URL `https://www.aclweb.org/anthology/N19-1423`.

Djellel Difallah, Elena Filatova, and Panos Ipeirotis. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 135–143, 2018.

Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G Ipeirotis, and Philippe Cudré-Mauroux. The dynamics of micro-task crowdsourcing: The case of amazon mturk. In *Proceedings of the 24th international conference on world wide web*, pages 238–247, 2015.

Souad Djelassi and Isabelle Decoopman. Customers' participation in product development through crowdsourcing: Issues and implications. *Industrial Marketing Management*, 42(5):683–692, 2013.

Anhai Doan, Raghu Ramakrishnan, and Alon Y Halevy. Crowdsourcing systems on the world-wide web. *Communications of the ACM*, 54(4):86–96, 2011.

Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1013–1022. ACM, 2012.

H Drucker, H Drucker, CJC Burges, L Kaufman, Kaufman BL C CJ, et al. Support vector regression machines advances in neural information processing systems 9. *Advances in Neural Information Processing Systems*, 9, 1996.

David Durward, Ivo Blohm, and Jan Marco Leimeister. Is there papa in crowd work?: A literature review on ethical dimensions in crowdsourcing. In *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*, pages 823–832. IEEE, 2016.

John E Edlund, Kathleene M Lange, Andrea M Sevene, Jonathan Umansky, Cassandra D Beck, and Daniel J Bell. Participant crosstalk: Issues when using the mechanical turk. *Tutorials in Quantitative Methods for Psychology*, 13:174–182, 2017.

Enrique Estellés-Arolas and Fernando González-Ladrón-de Guevara. Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200, 2012.

James A Euchner. The limits of crowds. *Research Technology Management*, 53 (5):7, 2010.

Wei Feng, Zheng Yan, Hengrun Zhang, Kai Zeng, Yu Xiao, and Y Thomas Hou. A survey on security, privacy, and trust in mobile crowdsourcing. *IEEE Internet of Things Journal*, 5(4):2971–2992, 2017.

Ailbhe Finnerty, Pavel Kucherbaev, Stefano Tranquillini, and Gregorio Convertino. Keep it simple: Reward and task design in crowdsourcing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*, pages 1–4, 2013.

George Fishman. *Monte Carlo: Concepts, algorithms, and applications*. Springer Science & Business Media, 2013.

Floyd Jackson Fowler Jr. How unclear terms affect survey data. *Public Opinion Quarterly*, 56(2):218–231, 1992.

Michael J Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, and Reynold Xin. Crowddb: Answering queries with crowdsourcing. In *Proceedings of the*

2011 ACM SIGMOD International Conference on Management of data, pages 61–72. ACM, 2011.

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

Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 5–14, 2017.

Snehalkumar (Neil) S Gaikwad, Mark E Whiting, Dilrukshi Gamage, Catherine A Mullings, Dinesh Majeti, Shirish Goyal, Aaron Gilbee, Nalin Chhibber, Adam Ginzberg, Angela Richmond-Fuller, et al. The daemo crowdsourcing marketplace. In *Companion of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1–4, 2017.

Mrunal Gawade, Rajan Vaish, Mercy Nduta Waihumbu, and James Davis. Exploring employment opportunities through microtasks via cybercafes. In *2012 IEEE Global Humanitarian Technology Conference*, pages 77–82. IEEE, 2012.

David Geiger, Stefan Seedorf, Thimo Schulze, Robert C Nickerson, and Martin Schader. Managing the crowd: Towards a taxonomy of crowdsourcing processes. In *AMCIS*, 2011.

Maarten Gelderman. Task difficulty, task variability and satisfaction with management support systems. *Information & Management*, 39(7):593–604, 2002.

Catherine Grady and Matthew Lease. Crowdsourcing document relevance assessment with mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical turk*, pages 172–179, 2010.

Mary L Gray, Siddharth Suri, Syed Shoaib Ali, and Deepti Kulkarni. The crowd is a collaborative network. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pages 134–147. ACM, 2016.

Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

Neha Gupta, David Martin, Benjamin V Hanrahan, and Jacki O'Neill. Turk-life in india. In *Proceedings of the 18th International Conference on Supporting Group Work*, pages 1–11. ACM, 2014.

Kristen L Guth and Daren C Brabham. Finding the diamond in the rough: Exploring communication and platform in crowdsourcing performance. *Communication Monographs*, 84(4):510–533, 2017.

Philipp Gutheim and Björn Hartmann. Fantasktic: Improving quality of results for novice crowdsourcing users. *EECS Dept., Univ. California, Berkeley, CA, USA, Tech. Rep. UCB/EECS-2012*, 112, 2012.

Buddhadeb Halder. Evolution of crowdsourcing: Potential data protection, privacy and security concerns under the new media age. *Revista Democracia Digital e Governo Eletrônico*, 1(10):377–393, 2014.

Christopher Harris. You're hired! an examination of crowdsourcing incentive models in human resource tasks. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 15–18, 2011a.

Christopher G Harris. Dirty deeds done dirt cheap: A darker side to crowdsourcing. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 1314–1317. IEEE, 2011b.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: Data mining, inference, and prediction*, volume 2. Springer, 2009.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543, 2021. URL `https://arxiv.org/abs/2111.09543`.

Mark Hedges and Stuart Dunn. *Academic crowdsourcing in the humanities: Crowds, communities and co-production*. Chandos Publishing, 2017.

Lars Hetmank. Components and functions of crowdsourcing systems–a systematic literature review. 2013.

Mokter Hossain and Ilkka Kauranen. Crowdsourcing: A comprehensive literature review. *Strategic Outsourcing: An International Journal*, 2015.

Mahmood Hosseini, Keith Phalp, Jacqui Taylor, and Raian Ali. The four pillars of crowdsourcing: A reference model. In *2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)*, pages 1–12. IEEE, 2014.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, 2013.

Jeff Howe. *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House, 2008.

Jeff Howe et al. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

Eric Huang, Haoqi Zhang, David C Parkes, Krzysztof Z Gajos, and Yiling Chen. Toward automatic task design: A progress report. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 77–85, 2010.

Shih-Wen Huang and Wai-Tat Fu. Don't hide in the crowd!: Increasing social transparency between peer workers improves crowdsourcing outcomes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 621–630. ACM, 2013.

Kosetsu Ikeda, Atsuyuki Morishima, Habibur Rahman, Senjuti Basu Roy, Saravanan Thirumuruganathan, Sihem Amer-Yahia, and Gautam Das. Collaborative crowdsourcing with crowd4u. *Proceedings of the VLDB Endowment*, 9(13): 1497–1500, 2016.

Panagiotis G Ipeirotis. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2):16–21, 2010.

Lilly C Irani and Max Six Silberman. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 611–620. ACM, 2013.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

Heehyoung Jang, Lorne Olfman, Ilsang Ko, Joon Koh, and Kyungtae Kim. The influence of on-line brand community characteristics on community commitment and brand loyalty. *International Journal of Electronic Commerce*, 12(3):57–80, 2008.

Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

Allen C Johnston, Merrill Warkentin, and Xin Luo. National culture and information privacy: The influential effects of individualism and collectivism on privacy concerns and organizational commitment. In *Proceedings of the International Federation of Information Processing (IFIP), International Workshop on Information Systems Security Research*, pages 88–104, 2009.

Prateek Joshi. Text mining 101: A stepwise introduction to topic modeling using latent semantic analysis (using python). *Analytics Vidhya, October*, 1, 2018.

Junegak Joung, Kiwook Jung, Sanghyun Ko, and Kwangsoo Kim. Customer complaints analysis using text mining and outcome-driven innovation method for market-oriented product development. *Sustainability*, 11(1):40, 2018.

Timothy A Judge and John D Kammeyer-Mueller. Job attitudes. *Annual review of psychology*, 63:341–367, 2012.

Daniel Jurafsky and James H Martin. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall, 2009.

Evgeny Kaganer, Erran Carmel, Rudy Hirschheim, and Timothy Olsen. Managing the human cloud. *MIT Sloan Management Review*, 54(2):23, 2013.

Arne L Kalleberg. Precarious work, insecure workers: Employment relations in transition. *American sociological review*, 74(1):1–22, 2009.

Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara Kiesler. Privacy attitudes of mechanical turk workers and the us public. In *10th Symposium On Usable Privacy and Security ({SOUPS} 2014)*, pages 37–49, 2014.

Gabriella Kazai. An exploration of the influence that task parameters have on the performance of crowds. *Proceedings of the CrowdConf*, 2010, 2010.

Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1941–1944. ACM, 2011.

Collins-Thompson Kevyn. Computational assessment of text readability. *ITL-International Journal of Applied Linguistics*, 165(2):97–135, 2014.

Shashank Khanna, Aishwarya Ratan, James Davis, and William Thies. Evaluating and improving the usability of mechanical turk for low-income workers in india. In *Proceedings of the first ACM symposium on computing for development*, pages 1–10, 2010.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.

Aniket Kittur. Crowdsourcing, collaboration and creativity. *XRDS: Crossroads, the ACM magazine for students*, 17(2):22–26, 2010.

Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456, 2008.

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

Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1301–1318, 2013.

Giordano Koch, Johann Füller, and Sabine Brunswicker. Online crowdsourcing in the public sector: How to design open government platforms. In *International Conference on Online Communities and Social Computing*, pages 203–212. Springer, 2011.

Anand Kulkarni, Philipp Gutheim, Prayag Narula, David Rolnitzky, Tapan Parikh, and Björn Hartmann. Mobileworks: Designing for quality in a managed crowdsourcing architecture. *IEEE Internet Computing*, 16(5):28–35, 2012.

Anand P Kulkarni, Matthew Can, and Bjoern Hartmann. Turkomatic: Automatic recursive task and workflow design for mechanical turk. In *CHI'11 extended abstracts on human factors in computing systems*, pages 2053–2058. 2011.

Karim R Lakhani and Robert G Wolf. Why hackers do what they do: Understanding motivation and effort in free/open source software projects. 2003.

Walter S Lasecki, Jaime Teevan, and Ece Kamar. The cost of asking crowd workers to behave maliciously. In *Proc. the AAMAS Workshop on Human-Agent Interaction Design and Models*, 2015.

Matthew Lease, Jessica Hullman, Jeffrey Bigham, Michael Bernstein, Juho Kim, Walter Lasecki, Saeideh Bakhshi, Tanushree Mitra, and Robert Miller. Mechanical turk is not anonymous. *Available at SSRN 2228728*, 2013.

Nedim Lipka and Benno Stein. Identifying featured articles in wikipedia: Writing style matters. In Michael Rappa, Paul Jones, Juliana Freire, and Soumen Chakrabarti, editors, *19th International Conference on World Wide Web (WWW 2010)*, pages 1147–1148. ACM, April 2010. ISBN 978-1-60558-799-8. doi:10.1145/1772690.1772847.

Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. Turkit: Human computation algorithms on mechanical turk. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 57–66, 2010.

An Liu, Weiqi Wang, Shuo Shang, Qing Li, and Xiangliang Zhang. Efficient task assignment in spatial crowdsourcing with worker and task privacy protection. *GeoInformatica*, 22(2):335–362, 2018.

Alessia Mammone, Marco Turchi, and Nello Cristianini. Support vector machines. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(3):283–289, 2009.

VK Chaithanya Manam and Alexander J Quinn. Wingit: Efficient refinement of unclear task instructions. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*, 2018.

VK Chaithanya Manam, Dwarakanath Jampani, Mariam Zaim, Meng-Han Wu, and Alexander J. Quinn. Taskmate: A mechanism to improve the quality of instructions in crowdsourcing. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1121–1130, 2019.

David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. Being a turker. In *Proc. 17th ACM Conf. Computer Supported Cooperative Work & Social Computing*, pages 224–235. ACM, 2014.

Winter Mason and Siddharth Suri. Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1):1–23, 2012.

Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. Taking a hit: Designing around rejection, mistrust, risk, and workers' experiences in amazon mechanical turk. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2271–2282, 2016.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 262–272. Association for Computational Linguistics, 2011.

Louise Muhdi, Michael Daiber, Sascha Friesike, and Roman Boutellier. The crowdsourcing process: An intermediary mediated idea generation approach in

the early phase of innovation. *International Journal of Entrepreneurship and Innovation Management*, 14(4):315–332, 2011.

Kevin P Murphy. *Machine learning: A probabilistic perspective*. MIT press, 2012.

Lobna Nassar and Fakhri Karray. Overview of the crowdsourcing process. *Knowledge and Information Systems*, 60(1):1–24, 2019.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics, 2010.

Zahra Nouri, Henning Wachsmuth, and Gregor Engels. Mining crowdsourcing problems from discussion forums of workers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6264–6276, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi:10.18653/v1/2020.coling-main.551. URL `https://www.aclweb.org/anthology/2020.coling-main.551`.

Zahra Nouri, Ujwal Gadiraju, Gregor Engels, and Henning Wachsmuth. What is unclear? computational assessment of task clarity in crowdsourcing. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 165–175, 2021a.

Zahra Nouri, Nikhil Prakash, Ujwal Gadiraju, and Henning Wachsmuth. iClarify–A tool to help requesters iteratively improve task descriptions in crowdsourcing. In *Proceedings of the 9th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2021b.

Zahra Nouri, Nikhil Prakash, Ujwal Gadiraju, and Henning Wachsmuth. Supporting requesters in writing clear crowdsourcing task descriptions through computational flaw assessment. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 737–749, 2023. doi:10.1145/3581641.3584039.

Debra Osnowitz and Kevin D Henson. Leveraging limits for contract professionals: Boundary work and control of working time. *Work and Occupations*, 43(3):326–360, 2016.

Jacki O'Neill and David Martin. Relationship-based business process crowdsourcing? In *IFIP Conference on Human-Computer Interaction*, pages 429–446. Springer, 2013.

Alexandra Papoutsaki, Hua Guo, Danae Metaxa-Kakavouli, Connor Gramazio, Jeff Rasley, Wenting Xie, Guan Wang, and Jeff Huang. Crowdsourcing from

scratch: A pragmatic experiment in data collection by novice requesters. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 3, 2015.

Jay Pedersen, David Kocsis, Abhishek Tripathi, Alvin Tarrell, Aruna Weerakoon, Nargess Tahmasbi, Jie Xiong, Wei Deng, Onook Oh, and Gert-Jan De Vreede. Conceptual foundations of crowdsourcing: A review of is research. In *2013 46th Hawaii International Conference on System Sciences*, pages 579–588. IEEE, 2013.

Xin Peng, Muhammad Ali Babar, and Christof Ebert. Collaborative software development platforms for crowdsourcing. *IEEE Software*, 31(2):30–36, 2014.

Isaac Persing and Vincent Ng. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269, 2013.

Constance Elise Porter. A typology of virtual communities: A multi-disciplinary foundation for future research. *Journal of Computer-mediated Communication*, 10(1):JCMC1011, 2004.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM, 2015.

Joel Ross, Lilly Irani, Max Six Silberman, Andrew Zaldivar, and Bill Tomlinson. Who are the crowdworkers? Shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*, pages 2863–2872. 2010.

Ann Marie Ryan and Jennifer L Wessel. Implications of a changing workforce and workplace for justice perceptions and expectations. *Human Resource Management Review*, 25(2):162–175, 2015.

Niloufar Salehi, Jaime Teevan, Shamsi Iqbal, and Ece Kamar. Communicating context to the crowd for complex writing tasks. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1890–1901, 2017.

Shruti Sannon and Dan Cosley. Privacy, power, and invisible labor on amazon mechanical turk. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

Gregory D Saxton, Onook Oh, and Rajiv Kishore. Rules of crowdsourcing: Models, issues, and systems of control. *Information Systems Management*, 30(1): 2–20, 2013.

Eric Schenk and Claude Guittard. Towards a characterization of crowdsourcing practices. *Journal of Innovation Economics & Management*, (1):93–107, 2011.

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

Thimo Schulze, Stefan Seedorf, David Geiger, Nicolas Kaufmann, and Martin Schader. Exploring task properties in crowdsourcing–an empirical study on mechanical turk. 2011.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

David Schwartz. Embedded in the crowd: Creative freelancers, crowdsourced work, and occupational community. *Work and Occupations*, 45(3), 2018.

Nada Sherief, Nan Jiang, Mahmood Hosseini, Keith Phalp, and Raian Ali. Crowdsourcing software evaluation. In *proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, page 19. ACM, 2014.

Jiangang Shu, Xiaohua Jia, Kan Yang, and Hua Wang. Privacy-preserving task recommendation services for crowdsourcing. *IEEE Transactions on Services Computing*, 2018.

Max Six Silberman. What's fair? Rational action and its residuals in an electronic market. *Unpublished manuscript*, 2010.

Max Six Silberman. *Human-centered computing and the future of work: Lessons from Mechanical Turk and Turkopticon, 2008-2015*. PhD thesis, UC Irvine, 2015.

Max Six Silberman, Lilly Irani, and Joel Ross. Ethics and tactics of professional crowdwork. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2):39–43, 2010a.

Max Six Silberman, Joel Ross, Lilly Irani, and Bill Tomlinson. Sellers' problems in human computation markets. In *Proceedings of the acm sigkdd workshop on human computation*, pages 18–21. ACM, 2010b.

Elena Simperl. How to use crowdsourcing effectively: Guidelines and examples. *Liber Quarterly*, 25(1):18–39, 2015.

Henri Simula. The rise and fall of crowdsourcing? In *System sciences (HICSS), 2013 46th Hawaii international conference on*, pages 2783–2791. IEEE, 2013.

Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. Learning from revisions: Quality assessment of claims in argumentation at scale. *arXiv preprint arXiv:2101.10250*, 2021.

Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14:199–222, 2004.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics, 2008.

Christian Stab and Iryna Gurevych. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, 2017.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961. Association for Computational Linguistics, 2012.

Tim Straub, Henner Gimpel, Florian Teschner, and Christof Weinhardt. Feedback and performance in crowd work: A real effort experiment. 2014.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Shaheen Syed and Marco Spruit. Selecting priors for latent dirichlet allocation. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pages 194–202. IEEE, 2018.

Klaus J Templer. Five-factor model of personality and job satisfaction: The importance of agreeableness in a tight and collectivistic asian society. *Applied Psychology*, 61(1):114–129, 2012.

Lois E Tetrick, Kelley J Slack, Nancy Da Silva, and Robert R Sinclair. A comparison of the stress–strain process for business owners and nonowners: Differences

in job demands, emotional exhaustion, satisfaction, and social support. *Journal of occupational health psychology*, 5(4):464, 2000.

Edmund R Thompson and Florence TT Phua. A brief index of affective job satisfaction. *Group & Organization Management*, 37(3):275–307, 2012.

Sheema Usmani, Mariana Bernagozzi, Yufeng Huang, Michelle Morales, Amir Sabet Sarvestani, and Biplav Srivastava. Clarity: Data-driven automatic assessment of product competitiveness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13204–13211, 2020.

Donna Vakharia and Matthew Lease. Beyond mechanical turk: An analysis of paid crowd work platforms. *Proceedings of the iConference*, pages 1–17, 2015.

Tanja Van der Lippe, Leonie Van Breeschoten, and Margriet Van Hek. Organizational work–life policies and the gender wage gap in european workplaces. *Work and Occupations*, 46(2):111–148, 2019. doi:10.1177/0730888418791652. URL `https://doi.org/10.1177/0730888418791652`.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

Lav R Varshney. Privacy and reliability in crowdsourcing service delivery. In *2012 Annual SRII Global Conference*, pages 55–60. IEEE, 2012.

Lav R Varshney, Aditya Vempaty, and Pramod K Varshney. Assuring privacy and reliability in crowdsourcing with coding. In *2014 Information Theory and Applications Workshop (ITA)*, pages 1–6. IEEE, 2014.

A Vaswani, N Shazeer, N Parmar, J Uszkoreit, et al. Attention is all you need in advances in neural information processing systems. *Search PubMed*, pages 5998–6008, 2017.

J Verschoore, Lucas Borella, and I Bortolaso. Towards a framework for crowdsourcing process management: Evidences from brazilian leading experts. *J. Bus. Econ.*, 6(1):187–203, 2015.

Henning Wachsmuth and Till Werner. Intrinsic quality assessment of arguments. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6739–6745, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi:10.18653/v1/2020.coling-main.592. URL `https://www.aclweb.org/anthology/2020.coling-main.592`.

Paul Wais, Shivaram Lingamneni, Duncan Cook, Jason Fennell, Benjamin Goldenberg, Daniel Lubarov, David Marin, and Hari Simons. Towards building a

high-quality workforce with mechanical turk. *Proceedings of computational social science and the wisdom of crowds (NIPS)*, pages 1–5, 2010.

Hanna M Wallach, David M Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In *Advances in neural information processing systems*, pages 1973–1981, 2009a.

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112. ACM, 2009b.

Jing Wang, Panagiotis G Ipeirotis, and Foster Provost. Managing crowdsourcing workers. In *The 2011 winter conference on business intelligence*, pages 10–12. Citeseer, 2011.

Daniel S Weld, Christopher H Lin, and Jonathan Bragg. Artificial intelligence and collective intelligence. *Handbook of Collective Intelligence*, pages 89–114, 2015.

Nanda Wijermans, Claudine Conrado, Maarten van Steen, Claudio Martella, and Jie Li. A landscape of crowd-management support: An integrative approach. *Safety science*, 86:142–164, 2016.

Kevin Wise, Brian Hamman, and Kjerstin Thorson. Moderation, response rate, and message interactivity: Features of online communities and their effects on intent to participate. *Journal of Computer-Mediated Communication*, 12(1):24–41, 2006.

Meng-Han Wu and Alexander James Quinn. Confusing the crowd: Task instruction quality on amazon mechanical turk. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*, 2017.

Huichuan Xia, Yang Wang, Yun Huang, and Anuj Shah. "Our privacy needs to be protected at all costs" crowd workers' privacy experiences on amazon mechanical turk. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW): 1–22, 2017.

Jie Yang, Carlo van der Valk, Tobias Hossfeld, Judith Redi, and Alessandro Bozzon. How do crowdworker communities and microtask markets influenceeach other? a data-driven study on amazon mechanical turk. Technical report, Université de Fribourg, 2018.

Cheng Zhang, Pingbo Tang, Pin-Chao Liao, and Yi Ren. Imagery-based risk assessment using crowdsourcing technology in complex workspaces. In *Computing in Civil Engineering 2017*, pages 174–182.

Yu Zhang and Mihaela van der Schaar. Reputation-based incentive protocols in crowdsourcing applications. In *INFOCOM, 2012 Proceedings IEEE*, pages 2140–2148. IEEE, 2012.

Yuxiang Zhao and Qinghua Zhu. Evaluation on crowdsourcing research: Current status and future direction. *Information Systems Frontiers*, 16(3):417–434, Jul 2014. ISSN 1572-9419. doi:10.1007/s10796-012-9350-4. URL `https://doi.org/10.1007/s10796-012-9350-4`.

Haichao Zheng, Dahui Li, and Wenhua Hou. Task design, motivation, and participation in crowdsourcing contests. *International Journal of Electronic Commerce*, 15(4):57–88, 2011.