Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Um estudo do uso de testes de qualificação na plataforma Amazon Mechanical Turk

Ianna Maria Sodré Ferreira de Sousa

Tese submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Doutor em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Redes de Computadores e Sistemas Distribuídos

Francisco Vilar Brasileiro
(Orientador)

Campina Grande, Paraíba, Brasil
©Ianna Maria Sodré Ferreira de Sousa, 19/07/2017

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

S725e Sousa, Ianna Maria Sodré Ferreira de.
Um estudo do uso de testes de qualificação na plataforma Amazon Mechanical Turk / Ianna Maria Sodré Ferreira de Sousa. – Campina Grande, 2017.
122 f. : il. color.

Tese (Doutorado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2017.
"Orientação: Prof. Dr. Francisco Vilar Brasileiro".
Referências.

1. Crowdsourcing. 2. Mercado de Microtarefas. 3. Mturk. 4. Qualificações. 5. Pré-seleção de Trabalhadores. I. Brasileiro, Francisco Vilar. II. Título.

CDU 004.822:005.953.2(043)

M ESTUDO DO USO DE TESTES DE QUALIFICAÇÃO NA PLATAFORMA AMAZON MECHANICAL TURK"

IANNA MARIA SODRÉ FERREIRA DE SOUSA

TESE APROVADA EM 19/07/2017

FRANCISCO VILAR BRASILEIRO, Ph.D, UFCG
Orientador(a)

NAZARENO FERREIRA DE ANDRADE, Dr., UFCG
Examinador(a)

DALTON DARIO SEREY GUERRERO, Dr., UFCG
Examinador(a)

RICARDO MATSUMURA DE ARAÚJO, D.Sc., UFPEL
Examinador(a)

JOÃO JOSÉ VASCO PEIXOTO FURTADO, PhD., UNIFOR
Examinador(a)

CAMPINA GRANDE - PB

Resumo

Muitos sistemas de computação por humanos usam mercados de trabalho *crowdsourcing* para recrutar trabalhadores. No entanto, devido à natureza aberta desses mercados, garantir que os resultados produzidos pelos trabalhadores possuam uma qualidade suficientemente alta ainda é uma tarefa desafiadora, particularmente em mercados de microtarefas, onde a avaliação precisa ser feita de forma automática. A pré-seleção de trabalhadores adequados é um mecanismo que pode melhorar a qualidade dos resultados obtidos. Isso pode ser feito considerando as informações do cadastro pessoal do trabalhador, o comportamento histórico do trabalhador no sistema ou o uso de testes de qualificação customizados. Entretanto, pouco se sabe sobre como os solicitantes usam testes de qualificação na prática e se estes tem influência na qualidade dos resultados apresentados pelos trabalhadores. Este estudo visa avançar esse conhecimento. Por meio de análise de distribuições, classificação e agrupamento, as tarefas e os solicitantes foram caracterizados utilizando dados obtidos da plataforma Amazon Mechanical Turk em dois períodos de tempo distintos. Os resultados mostram que a maioria das tarefas (94% e 87%, para a coleta de dados 1 e 2, respectivamente) usa algum teste de qualificação para a pré-seleção de trabalhadores e que o tipo e o número de testes de qualificação não são determinados pela classe da tarefa. Os solicitantes, em sua maioria, submetem tarefas com apenas um único teste de qualificação do tipo reputação, no entanto, os solicitantes mais ativos na plataforma usam, exclusivamente, teste de qualificação customizado. Para avaliar o impacto do uso de testes de qualificação customizados na qualidade dos resultados produzidos, foram realizados experimentos com três tipos diferentes de tarefas usando tanto trabalhadores qualificados (mestres ou trabalhadores pré-selecionados) como não qualificados. Os resultados mostram que a pontuação média alcançada pelos trabalhadores pré-selecionados foi sempre maior que a alcançada por trabalhadores que não foram pré-selecionados. Além disso, o desempenho de trabalhadores pré-selecionados foi muito próximo dos trabalhadores considerados mestres e, em alguns cenários, melhor, indicando assim, que é possível obter resultados mais acurados em plataformas de trabalho on-line de microtarefas quando se usa testes de qualificação.

Abstract

Many human computation systems use crowdsourcing labor markets to recruit workers. However, it is still a challenge to guarantee that the results produced by workers have a high enough quality. This is particularly difficult in markets based on micro-tasks, where the assessment of the quality of the results needs to be done automatically. Pre-selection of suitable workers is a mechanism that can improve the quality of the results achieved. This can be done by considering worker's personal information, worker's historical behavior in the system, or the use of customized qualification tasks. However, little is known about how requesters use qualification tests in practice. This study aims at advance this knowledge. Through the analysis of distributions, classification and grouping, the tasks and the requesters were characterized using data obtained from the Amazon Mechanical Turk platform in two different time periods. Results show that most jobs (94% and 87% for the two collection periods, respectively) use some qualification test for the pre-selection of workers and that the type and number of qualification tests are not determined by the task class. Requesters, for the most part, submit tasks with only a single reputation-type qualification test, however, the most active requesters on the platform use only customized qualification tests. To assess the impact that the use of customized qualifications has in the quality of the results produced, we have executed experiments with three different types of tasks using both unqualified and qualified workers (masters and pre-selected workers). The results showed that, generally, qualified workers provide more accurate answers, when compared to unqualified ones. In addition, the performance of pre-selected workers was very close to the workers considered masters and, in some scenarios, it was better. Finally, it is possible to obtain more accurate results in micro-task online work platforms when using qualification tests.

Agradecimentos

Aos colegas do Laboratório de Sistemas Distribuídos, que contribuíram, de uma forma ou de outra, para a conclusão desse trabalho. Dentre eles sou grata a David Candeia, Adabriand Furtado e, especialmente, a Lesandro Ponciano, pelo incentivo na realização desse trabalho e pelas discussões sobre o conteúdo do mesmo.

Minha gratidão aos professores Dalton Guerrero, João José Vasco Peixoto Furtado, Nazareno Andrade e Ricardo Matsumura pelos importantes comentários feitos na defesa da qualificação e que contribuíram para a condução da pesquisa e finalização deste documento.

Meu agradecimento especial ao professor Fubica pelos conselhos, paciência, confiança e suporte em todas as etapas desse trabalho. Foi uma honra e orgulho tê-lo como orientador.

Finalmente, agradecer a presença amorosa, a ajuda e o estímulo de meu esposo e dos meus filhos, Felipe, Fabrício, Marianna e Marinna. Sem o amor e a compreensão deles, esse objetivo não teria sido alcançado.

Conteúdo

1 Introdução **1**

1.1 Problema 3

1.2 Objetivos 5

1.3 Resultados e contribuições 7

1.4 Estrutura do documento 8

2 Computação por Humanos **10**

2.1 Definição e conceitos relacionados 11

2.2 Sistemas de computação por humanos 14

2.3 *Crowdsourcing* remunerado 16

2.4 Trabalhadores, solicitantes e motivação 19

2.5 Considerações finais 21

3 A qualidade dos resultados em mercados de trabalho on-line e a plataforma MTurk **23**

3.1 Estratégias de controle de qualidade dos resultados 25

3.1.1 Controle de qualidade após submissão de resultados 25

3.1.2 Controle de qualidade preventivo 29

3.2 Mechanical Turk 33

3.2.1 Funcionamento do MTurk 35

3.2.2 Testes de Qualificação no MTurk 37

3.3 Considerações finais 43

4 Uso de qualificação no MTurk: pesquisa exploratória quantitativa **45**

4.1 Materiais e métodos 46

4.1.1 Coleta de dados 46
4.1.2 Classificação das tarefas 48
4.1.3 Classificação dos solicitantes 52
4.2 Apresentação e análise dos resultados 53
4.2.1 Conjuntos de dados coletados da plataforma MTurk 53
4.2.2 Os solicitantes 53
4.2.3 Classificação das tarefas 55
4.2.4 Distribuição do uso de testes de qualificação 62
4.2.5 Como os solicitantes usam as qualificações 72
4.2.6 Distribuição da recompensa 84
4.3 Considerações finais 88

5 O efeito do teste de qualificação customizado na qualidade dos resultados 89
5.1 Materiais e métodos 90
5.1.1 Descrição dos experimentos 90
5.1.2 Testes de qualificação 92
5.2 Apresentação e análise dos resultados 92
5.3 Considerações finais 100

6 Conclusões 101
6.1 Resultados e contribuições 101
6.2 Limitações 105
6.3 Trabalhos futuros 106

A Tarefas utilizadas nos experimentos 119

Lista de Siglas

API Acrônimo da expressão *Application Programming Interface*, geralmente traduzida como Interface de Programação de Aplicações.

BORW Acrônimo da expressão *Bag-Of-Related-Words*, abordagem para representação de coleção de documentos.

HIT Acrônimo da expressão *Human Intelligence Task*, traduzida como Tarefas que Requerem Inteligência Humana. É sinônimo de tarefas de computação por humanos.

MTurk Acrônimo de *Mechanical Turk*. Trata-se de um sistema de computação por humanos de propriedade da empresa Amazon.com, Inc.

NASA Acrônimo da expressão *National Aeronautics and Space Administration*, geralmente traduzida como Administração Nacional da Aeronáutica e do Espaço. Trata-se de um organização do governo dos Estados Unidos.

OCR Acrônimo da expressão *Optical Character Recognition* traduzida como Reconhecimento Ótico de Caracteres.

reCAPTCHA Um serviço utilizado para proteger sítios Web de ataques de *spammers* automatizados. O serviço é baseado no conceito de computação por humanos.

SVM Acrônimo da expressão *Support Vector Machine*, em geral traduzida como Máquina de Vetores de Suporte.

URL Endereço de um recurso disponível em uma rede, seja a rede internet ou intranet, e significa em inglês *Uniform Resource Locator*.

Lista de Figuras

2.1 Diagrama de Venn para ilustrar como os termos computação por humanos, *crowdsourcing* e computação social se relacionam. Adaptado de Quinn e Bederson (2011) 13

2.2 Esquema de funcionamento de plataforma *crowdsourcing*. Adaptado de Hirth et al.(2013). 19

3.1 Esquema de funcionamento da votação majoritária. Adaptado de Hirth et al. (2013). 27

3.2 Esquema de funcionamento da estratégia de votação majoritária com grupo de controle. Adaptado de Hirth et al. (2013). 28

3.3 Quadro de tarefas do MTurk utilizado pelos trabalhadores para seleção de tarefas. Fonte: www.mturk.com 36

3.4 Interface para criar tarefas e escolher teste de qualificação no Mturk. Fonte: www.mturk.com. 42

3.5 Quadro de testes de qualificação do tipo habilidade no MTurk. Fonte: www.mturk.com. 43

4.1 Perfil dos solicitantes considerando o número de tarefas submetidas 54

4.2 Número de termos que aparecem em pelo menos x% das tarefas 56

4.3 Distribuição das tarefas nas classes consideradas para o conjunto de dados I. 60

4.4 Distribuição das tarefas nas classes consideradas para o conjunto de dados II. 60

4.5 Porcentagem de tarefas que usam um número particular de qualificações no conjunto de dados I. 63

4.6 Porcentagem de tarefas que usam um número particular de qualificações no conjunto de dados II. 63

4.7 Distribuição do número de qualificações solicitadas nas diferentes classes de tarefas do conjunto de dados I. 68
4.8 Distribuição do número de qualificações solicitadas nas diferentes classes de tarefas do conjunto de dados II. 69
4.9 Distribuição do uso de diferentes tipos de qualificação nas diferentes classes de tarefas do conjunto de dados I. 70
4.10 Distribuição do uso de diferentes tipos de qualificação nas diferentes classes de tarefas do conjunto de dados II. 71
4.11 Distribuição do número de qualificações exigidas nos diferentes grupos de solicitantes do conjunto de dados I. 75
4.12 Distribuição do número de qualificações exigidas nos diferentes grupos de solicitantes do conjunto de dados II. 76
4.13 Distribuição do uso de 1 e 2 qualificações para tarefas das classes S e N considerando os grupos G2 e G4 do conjunto de dados I. 77
4.14 Distribuição do uso de 1 e 2 qualificações para tarefas das classes S e N considerando os grupos G2 e G4 do conjunto de dados II. 78
4.15 Distribuição do tipo de qualificação exigida nos diferentes grupos de solicitantes do conjunto de dados I. 79
4.16 Distribuição do tipo de qualificação exigida nos diferentes grupos de solicitantes do conjunto de dados II. 80
4.17 Distribuição de 1 e 2 exigências de qualificações do tipo Reputação e Padronizada para cada grupo de solicitantes do conjunto de dados I. 81
4.18 Distribuição de 1 e 2 exigências de qualificações do tipo Reputação e Padronizada para cada grupo de solicitantes do conjunto de dados II. 82
4.19 Distribuição da recompensa dos conjuntos de dados I e II. 84
4.20 Diferenças entre as médias das recompenses oferecidas pelos pares distintos de classes de tarefas do conjunto de dados I. 87
4.21 Diferenças entre as médias das recompensas oferecidas pelos pares distintos de classes de tarefas do conjunto de dados II. 87

5.1 Distribuição da pontuação para os testes de qualificação. 95

5.2 Distribuição da pontuação dos trabalhadores nos experimentos. 97

A.1 Tarefas da classe *Information Find* (IF) submetidas no MTurk. 120

A.2 Tarefas da classe *Interpretation and Analysis* (IA) submetidas no MTurk. . 121

A.3 Tarefas da classe *Content Crestion* (CC) submetidas no MTurk. 122

Lista de Tabelas

4.1 Classes e Subclasses da taxonomia utilizada para tarefas típicas em plataformas *crowdsourcing* 49
4.2 Resultado das coletas de dados. 53
4.3 Exemplo da etapa de pré-processamento dos campos das tarefas. 55
4.4 Termos relevantes (unigramas e bigramas) usados na classificação e o suporte associado do Conjunto de Dados I. 57
4.5 Número de tarefas no conjunto de treino e de teste de cada conjunto de dados. 58
4.6 Matriz confusão do modelo SVM para o conjunto de dados I. 59
4.7 Matriz confusão do modelo SVM para o conjunto de dados II. 59
4.8 Resumo da distribuição dos testes de qualificação utilizados nas tarefas do conjunto de dados I. 64
4.9 Resumo da distribuição dos testes de qualificação usados nas tarefas do conjunto de dados II. 65
4.10 Descrição dos grupos de solicitantes do conjunto de dados I. 73
4.11 Descrição dos grupos de solicitantes do conjunto de dados II. 73

5.1 Testes de qualificações do tipo habilidade utilizados nos experimentos . . . 92
5.2 Tabela das ações escolhidas para os experimentos 93
5.3 Descrição das tarefas utilizadas nos experimentos. 94
5.4 Tipos de qualificações customizadas criadas para os experimentos. 94
5.5 Intervalo de confiança para a média da acurácia da população 99

Capítulo 1

Introdução

Antes da primeira metade do século XX, o termo computador era usado para designar seres humanos que usavam suas habilidades cognitivas para realizar cálculos matemáticos complexos (Grier, 2013). Depois esses computadores humanos foram substituídos por computadores digitais que foram capazes de executar os mesmos cálculos de uma maneira muito mais eficiente. No entanto, ainda há uma série de tarefas para as quais a capacidade cognitiva dos seres humanos excede em muito a capacidade dos computadores atuais de executá-las. Exemplos de tais tarefas são a compreensão de idéias, a expressão de opiniões, a identificação de objetos e outras atividades relacionadas à criatividade.

Para fazer uso do poder cognitivo humano para realizar tarefas em larga escala, um novo paradigma de computação surgiu, denominado Computação Humana Distribuída ou, simplesmente, Computação por Humanos. A ideia da computação por humanos é que seres humanos trabalhem em conjunto com os computadores, para juntos, solucionar problemas que nenhum dos dois consegue resolver eficientemente de forma isolada (Quinn and Bederson, 2011).

Em Ciência da Computação, em geral, o ser humano faz uma descrição formal do problema e recebe a solução do computador para ser analisada. Na computação por humanos, o ser humano, ou um grupo de seres humanos, resolve o problema que foi delegado pelo computador. Problema esse que é fácil para um ser humano, mas difícil para o computador. Assim, uma parte da computação é realizada por computadores e outra parte é realizada por

seres humanos.

Um sistema de computação por humanos pode ser definido como um sistema que organiza e agrega o poder cognitivo do ser humano, chamado de trabalhador, com o objetivo de encontrar uma solução precisa para problemas que exigem inteligência humana de forma mais eficiente (Law, 2011).

Com o crescimento da web, sistemas de computação por humanos podem aproveitar as habilidades de um número sem precedentes de pessoas através da web para realizar computação complexa. Para viabilizar o paradigma da computação por humanos é preciso utilizar uma estratégia para recrutar um grande número de pessoas para atuarem como processadores humanos. As estratégias mais utilizadas são os sistemas que implementam jogos com propósito, sistemas de pensamento distribuído e sistemas de trabalho on-line.

Nos sistemas que implementam jogos com propósito, os trabalhadores contribuem para a solução de um problema de forma implícita enquanto se divertem jogando. O exemplo mais conhecido é o jogo ESP (Von Ahn and Dabbish, 2008) que recolheu milhões de etiquetas de imagem e depois foi adotado pelo Google (*Google Imagem Labeler*) para ajudar a melhorar a pesquisa de imagens. Assim, a execução da tarefa não é o objetivo do trabalhador. Isso também acontece no sistema reCAPTCHA (Von Ahn et al., 2008), onde as pessoas geram conteúdo enquanto validam o acesso a uma área restrita de uma página web.

Nos sistemas de pensamento distribuído os trabalhadores realizam as tarefas como um trabalho voluntário, sem esperar nada em troca. A motivação do trabalhador está relacionada ao prazer de colaborar e de aprender. O exemplo mais conhecido é o Zooniverse[1], plataforma que já permitiu a execução de dezenas de milhões de tarefas (Sauermann and Franzoni, 2015).

Os sistemas de *crowdsourcing* são um tipo particular de sistema de computação por humanos que foram desenvolvidos para fornecer suporte aos participantes para encontrar tarefas e resolvê-las. Em sistemas de *crowdsourcing* pagos, participantes ou trabalhadores, como

[1] www.zooniverse.org

são comumente chamados, recebem recompensa financeira para executar as tarefas. Esses sistemas são chamados mercados de trabalho on-line. Um dos principais exemplos é a plataforma Amazon Mechanical Turk[2] (MturK) que recebe por dia entre 25.000 e 35.000 tarefas[3] para serem executadas e possui mais de 400 mil trabalhadores registrados (Ipeirotis, 2010b; Ross et al., 2010; Agrawal and Srikant, 1994). CrowdFlower, Elance e TopCoder também são exemplos de mercados de trabalho on-line. O mercado fornece recursos que permitem que os solicitantes (indivíduos, grupos ou organizações) publiquem tarefas e os trabalhadores encontrem tarefas apropriadas a serem resolvidas (Ponciano et al., 2014). O mercado também intermedia a compensação financeira oferecida pelos solicitantes aos trabalhadores para as tarefas que são concluídas com êxito, conforme julgamento dos solicitantes.

1.1 Problema

Os sistemas de trabalho on-line de microtarefas, em particular, possuem a característica de que as tarefas são pequenas e simples. Essas tarefas são definidas pelo solicitante (indivíduos, grupos ou organizações), que tem um problema a ser resolvido, e enviadas para o sistema. Os trabalhadores usam o sistema para procurar tarefas para realizar. Como as tarefas são simples os trabalhadores não precisam ser especialistas em um domínio específico, ou terem familiaridade com a tarefa antes de executá-la, possibilitando que praticamente qualquer pessoa possa participar do sistema. Os trabalhadores são, portanto, anônimos, e características como por exemplo experiência, competência, intenção e interesse, nem sempre são observadas.

A redução da barreira de entrada para os trabalhadores aumenta o número de trabalhadores potenciais que podem ingressar no mercado. No entanto, também aumenta a heterogeneidade na qualidade dos resultados produzidos por uma grande parcela de trabalhadores. A qualidade dos resultados é uma questão importante em tais mercados, porque os solicitantes, em geral, enviam centenas ou mesmo milhares de pequenas tarefas relacionadas ao mesmo tempo. Avaliar a qualidade do resultado de cada tarefa individualmente, antes de decidir se deve ou não realizar o pagamento ofertado ao trabalhador, portanto, não é uma tarefa

[2] www.mturk.com

[3] Sistema Mturk Tracker - http://www.mturk-tracker.com/#/arrivals, último acesso em 22 de maio de 2016

viável de ser realizada pelo solicitante. Por isso, os solicitantes precisam fazer uso de mecanismos que garantam a qualidade dos resultados apresentados pelos trabalhadores. Assim, eles precisam contar com meios automáticos para inferir probabilisticamente a qualidade dos resultados apresentados pelos trabalhadores.

Trabalhadores que têm o único objetivo de aumentar seu rendimento em geral submetem resultados de baixa qualidade porque procuram realizar as tarefas no menor tempo possível, muitas vezes em detrimento da qualidade do seu trabalho (Wang et al., 2013). Há também os trabalhadores maliciosos que tentam tirar vantagem do sistema de forma intencional submetendo resultados de baixa qualidade, na esperança de não serem pegos pela avaliação dos solicitantes. Além disso, há a situação em que trabalhadores legítimos, mas que não têm a habilidade necessária para executar uma determinada tarefa, também gerem resultados de baixa qualidade. Estudos mostram que 30% ou mais dos resultados no MTurk, a plataforma mais popular de microtarefas, pode ser de baixa qualidade (Kittur et al., 2008; Bernstein et al., 2015).

Como as plataformas de trabalho on-line não fornecem mecanismos rigorosos de controle e monitoramento da qualidade, a responsabilidade de selecionar trabalhadores e avaliar os resultados é dos solicitantes. Assim, um importante desafio neste tipo de mercado on-line é garantir que a qualidade dos resultados esteja acima de um determinado limiar de qualidade.

Várias alternativas foram criadas para melhorar a qualidade dos resultados em mercados de trabalho on-line pagos orientados para microtarefas. Algumas abordagens dizem respeito à garantia de que as tarefas são, por design, resistentes a trabalhadores de baixa qualidade. Outros se concentram em avaliar automaticamente os resultados produzidos, rejeitando os de baixa qualidade. Alguns trabalhos avaliaram a eficácia dessas abordagens (Kittur et al., 2008; Difallah et al., 2012; Khanna et al., 2010; Ipeirotis et al., 2010). As abordagens que avaliam os resultados após a submissão são mais adequadas para o caso de tarefas objetivas (Dow et al., 2011). Além disso, não há consenso de que possam ser generalizadas para qualquer tipo de tarefa.

Outra estratégia para tentar melhorar a qualidade dos resultados é utilizando um filtro, isto é, fazendo uma pré-seleção dos trabalhadores através da avaliação de habilidades necessárias para executar a tarefa. Essa estratégia pode ser utilizada de forma diferente pelas diversas plataformas. Por exemplo, a plataforma Crowdflower[4] utiliza um conjunto de perguntas com respostas conhecidas para testar a habilidade dos trabalhadores (Vakharia and Lease, 2013). No Mechanical Turk, os trabalhadores podem ser avaliados de acordo com sua reputação no sistema, capacidade e habilidade de realizar as tarefas. Os solicitantes podem adicionar requisitos de qualificação às tarefas que eles submetem, como uma tentativa de estabelecer critérios mínimos de aceitação de desempenho para os trabalhadores que executam essas tarefas. Considerando a percepção dos trabalhadores sobre os mecanismos de pré-seleção, esta é a alternativa que tem a melhor reputação entre os trabalhadores (Schulze et al., 2013). Além disso, existem indicações de que o uso de tal mecanismo melhora os resultados apresentados (Su et al., 2007), e ao mesmo tempo pode ser usado para desencorajar trabalhadores mal-intencionados (Zhu and Carterette, 2010).

Embora existam evidências de que os solicitantes usam os requisitos de qualificação, não há um estudo abrangente sobre como os solicitantes usam esse mecanismo e qual é o efeito desse mecanismo na qualidade dos resultados. Este trabalho visa preencher essa lacuna.

Diante do exposto, o problema tratado neste trabalho é o pouco conhecimento sobre a eficiência do uso de estratégia de pré-seleção de trabalhadores em sistemas de trabalho on-line e de como esse conhecimento pode ser útil para a definição das tarefas pelos solicitantes. Como estudo de caso, investiga-se o uso da estratégia em um sistema de trabalho on-line de microtarefa.

1.2 Objetivos

O principal objetivo deste trabalho é investigar se o uso de teste de qualificação do tipo customizado em tarefas submetidas em mercados de trabalho on-line de microtarefas tem influência na qualidade dos resultados. Para atingir esse objetivo geral, os seguintes objetivos

[4] www.crowdflower.com

específicos foram definidos:

1. Fazer um levantamento das estratégias existentes para controlar a qualidade dos resultados em plataformas *crowdsourcing* de microtarefas.

2. Investigar como os testes de qualificação são utilizados pelos solicitantes, para pré-selecionar trabalhadores para suas tarefas, na plataforma Amazon Mechanical Turk.

3. Verificar a influência do uso de teste de qualificação, do tipo customizado, na qualidade dos resultados submetidos pelos trabalhadores na plataforma Amazon Mechanical Turk.

O desenvolvimento desse trabalho iniciou-se com uma pesquisa exploratória na plataforma Amazon Mechanical Turk para identificar se os usuários, isto é, os solicitantes, utilizavam algum mecanismo em suas tarefas para controlar a qualidade dos resultados submetidos pelos trabalhadores e que mecanismos eram esses. Os resultados mostraram que a maioria dos usuários utiliza mecanismo para pré-selecionar seus trabalhadores. Os mecanismos identificados foram classificados como testes de qualificação do tipo reputação, padronizado e customizado. Em relação aos trabalhadores da plataforma, identificou-se a existência de trabalhadores denominados mestres, que são trabalhadores que possuem alta reputação no sistema. Assim sendo, as hipóteses a serem investigadas são:

- Os resultados submetidos por qualquer trabalhador e por trabalhador pré-selecionado são diferentes em termos de qualidade dos resultados submetidos. Dado que há uma grande diversidade de tarefas que pode ser realizada por qualquer trabalhador cadastrado na plataforma e que o trabalhador não precisa ter conhecimento específico para realizar as tarefas, espera-se que ao usar um mecanismo para pré-selecionar trabalhadores considerando as necessidades específicas da tarefa, este trabalhador apresente melhores resultados.

- Os resultados submetidos por trabalhadores pré-selecionados e por trabalhadores mestres são diferentes em termos de qualidade dos resultados submetidos. A possibilidade do solicitante usar trabalhadores mestres em suas tarefas leva a crer que o mesmo obterá resultados superiores aos submetidos por trabalhadores comuns. O que se deseja

é descobrir a relação entre os resultados de trabalhadores mestres e os resultados de trabalhadores pré-selecionados.

1.3 Resultados e contribuições

Os principais resultados e contribuições obtidos na pesquisa descrita neste trabalho são os seguintes:

- Propõe-se analisar as estratégias existentes para controlar a qualidade dos resultados em plataformas *crowdsourcing* de microtarefas com o objetivo de identificar as vantagens e desvantagens de cada uma delas. Mostra-se que as estratégias apresentadas são mais eficientes para tarefas objetivas e, que, apesar das estratégias de pré-seleção de trabalhadores exigirem mais esforço por parte dos solicitantes, parecem ser mais adequadas para o controle da qualidade dos resultados dos trabalhadores (Sodré and Brasileiro, 2017). Além disso, são estratégias conhecidas e aceitas pelos trabalhadores.

- A investigação de como os testes de qualificação são utilizados na plataforma Mechanical Turk foi realizada através de pesquisa exploratória quantitativa sobre dados coletados da plataforma buscando responder, também, o quanto são utilizados (Sodré and Brasileiro, 2017). Mostra-se que o número de testes de qualificação utilizado nas tarefas apresenta variação assim como o tipo de teste. As tarefas foram classificadas com o intuito de investigar se o tipo de tarefa apresentava influência no uso de testes de qualificação. Os resultados mostram que há uma variação considerável na distribuição de testes de qualificações nas diferentes classes de tarefas e que as tarefas podem variar em relação ao número de qualificações utilizadas e em relação ao tipo de qualificação na mesma classe. Logo, o tipo e o número de testes de qualificação utilizados nas tarefas não são determinados pela classe da tarefa. Em relação aos solicitantes tem-se que há muitas tarefas que são submetidas por poucos solicitantes e poucas tarefas que são submetidas por muitos solicitantes. Os solicitantes foram agrupados considerando a similaridade entre suas tarefas. Os resultados mostram que a maioria dos solicitantes utiliza apenas um único teste de qualificação do tipo reputação nas tarefas que submetem, enquanto que os solicitantes mais ativos na plataforma usam, exclusivamente,

teste de qualificação customizado. Isso pode ser uma indicação de que os tipos de testes de qualificação existentes na plataforma não sejam suficientes para fornecer o nível de filtragem exigido pelos solicitantes que investem mais pesadamente na plataforma. As análises também mostram que as tarefas que fazem a pré-seleção do trabalhador oferecem recompensas maiores do que as tarefas que não fazem e que, considerando a classificação de tarefas, existem classes que oferecem recompensas melhores do que outras.

- Propõe-se verificar a influência do uso de teste de qualificação, do tipo customizado, na qualidade dos resultados submetidos pelos trabalhadores. Experimentos foram realizados na plataforma Mechanical Turk utilizando o teste de qualificação do tipo customizado em tarefas construídas a partir de tarefas existentes na plataforma (Sodré and Brasileiro, 2017). Foram consideradas três classes de trabalhadores nos experimentos: trabalhadores mestres, trabalhadores pré-selecionados e todos os trabalhadores da plataforma. Os resultados mostraram que a pontuação média alcançada pelos trabalhadores pré-selecionados foi sempre maior que a alcançada por trabalhadores que não foram pré-selecionados. Além disso, o desempenho de trabalhadores pré-selecionados foi muito próximo dos trabalhadores considerados mestres e, em alguns cenários, melhor.

Em resumo, a pesquisa realizada amplia o entendimento sobre o uso de testes de qualificação para pré-selecionar trabalhadores e mostra que é possível obter resultados de melhor qualidade em plataformas de trabalho on-line de microtarefas quando se usa testes de qualificação.

1.4 Estrutura do documento

A organização deste trabalho está sistematizada da seguinte forma. O Capitulo 2 trata da contextualização e escopo do trabalho apresentando conceitos relacionados na área de conhecimento da pesquisa, além de apresentar os sistemas de computação por humanos, os participantes e os fatores motivacionais que podem impulsionar a atividade dos participantes em tais sistemas.

O Capítulo 3 apresenta a fundamentação teórica das estratégias de controle de qualidade existentes utilizadas para controlar a qualidade dos resultados em plataformas *crowdsourcing*, bem como a descrição do funcionamento da plataforma Mechanical Turk.

O Capítulo 4 apresenta os resultados de pesquisa exploratória quantitativa realizada para ampliar o entendimento de como os solicitantes utilizam testes de qualificação para pré-selecionar trabalhadores em mercado de trabalho on-line de microtarefas.

O capítulo 5 descreve o estudo realizado para avaliar o efeito do uso de testes de qualificação na qualidade dos resultados submetidos pelos trabalhadores em mercados de trabalho on-line.

O Capítulo 6 apresenta os resultados e contribuições da pesquisa descrita neste documento, suas limitações e possíveis perspectivas de pesquisas futuras.

Capítulo 2

Computação por Humanos

Computação por humanos é uma área de pesquisa recente e em evolução, que tem como objetivo aproveitar a inteligência humana para resolver problemas computacionais que ainda não são resolvidos de forma eficiente pelos algoritmos executados por processadores, apesar do poder computacional existente atualmente.

Os problemas que se enquadram nesta categoria são, em geral, problemas de Inteligência Artificial que incluem tarefas perceptivas (por exemplo, reconhecimento de objetos e classificação de música), processamento de linguagem natural (por exemplo, análise de sentimentos e tradução) e tarefas cognitivas complexas (por exemplo, planejamento e criatividade) (Law and Ahn, 2011). Mas, pesquisas também são desenvolvidas em outras áreas, como por exemplo, criptografia (Ahn, 2005), algoritmos genéticos (Kosorukoff, 2001) e interação homem-máquina (Bernstein et al., 2015; Hu et al., 2010; Bigham et al., 2010) e negócios (Sheng et al., 2008; Malone et al., 2009; Ipeirotis et al., 2010; Wolfers and Zitzewitz, 2004; Little et al., 2010).

O recrutamento das pessoas pode ser realizado de diferentes formas, isto é, através de jogos com um propósito (Von Ahn and Dabbish, 2008) ou através de trabalho implícito (Von Ahn et al., 2008). No entanto, grande parte das aplicações de computação por humanos utilizam *crowdsourcing* para recrutar colaboradores. Muitas pesquisas em computação por humanos, inclusive, são atribuídas ao termo *crowdsourcing*. *Crowdsourcing* é um tipo de atividade on-line participativa em que um indivíduo, organização ou empresa propõe a um

grupo de indivíduos, de conhecimentos variados, através de uma chamada aberta, o compromisso de realizar uma tarefa (Quinn and Bederson, 2011; Howe, 2006; Estellés-Arolas and González-Ladrón-de Guevara, Estellés-Arolas and González-Ladrón-de Guevara). Porém, nem todas as atividades realizadas pelos seres humanos em *crowdsourcing* necessitam de habilidades cognitivas. Assim, nem toda atividade definida como *crowdsourcing* é considerada também uma atividade de computação por humanos.

No entanto, além de recrutar as pessoas é preciso mantê-las engajadas ao trabalho, isto é, motivá-las a permanecer trabalhando e apresentando resultados de qualidade. Essas questões são discutidas no decorrer deste capítulo que apresenta a definição de computação por humanos e de termos relacionados, sistemas de computação por humanos, os participantes do sistema e os fundamentos da teoria da motivação.

2.1 Definição e conceitos relacionados

O termo computação por humanos foi popularizado na área de ciência da computação em 2005, com a tese de doutorado de Luis von Ahn, intitulada *"Human Computation"*, que define o termo como "um paradigma para a utilização do poder de processamento humano para resolver os problemas que os computadores ainda não podem resolver"(Ahn, 2005).

Fazendo uma analogia com a computação tradicional, nessa definição, o ser humano, mais especificamente, seu cérebro, atua como um processador. Dessa forma, a diferença entre a computação tradicional e a computação por humanos está relacionada ao processador utilizado. Na computação tradicional uma pessoa utiliza o computador para resolver problemas utilizando algoritmos computacionais e, na computação por humanos o computador faz uma pessoa, ou grupo de pessoas, resolver problemas, agregando, posteriormente, os resultados e obtendo a solução final.

Outras definições encontradas na literatura sugerem explicitamente o envolvimento de muitas pessoas, ou seja, uma multidão (*crowd*), nos sistemas de computação por humanos. Quinn e Bederson (2011), por exemplo, definem computação por humanos como *"sistemas de computadores e de um grande número de seres humanos que trabalham em conjunto a*

fim de resolver problemas que não podem ser resolvidos por computadores ou por humanos isoladamente", enquanto que Law et al. (2009) como sendo *"uma nova área de pesquisa que estuda o processo de canalização da vasta população da Internet para realizar tarefas ou fornecer dados para resolver problemas difíceis que nenhum algoritmo computacional eficiente pode ainda resolver"*.

Outros trabalhos também apresentam taxonomia para computação por humanos, porém utilizando critérios diferentes para a classificação dos sistemas. Os sistemas são classificados, por exemplo, com base no objetivo (projeto e inovação, desenvolvimento e teste, marketing, vendas e suporte, são exemplos), na forma de recrutamento (por exemplo, competição e mercado de trabalho), algoritmos e bancos de dados (Yuen et al., 2009; Vukovic, 2009)

A partir das definições pode-se perceber que há um consenso de que os problemas da área de computação por humanos se encaixam no paradigma da computação em geral, e como tal um dia poderão ser resolvidos por computadores (Quinn and Bederson, 2011).

Na literatura existem termos relacionados à computação por humanos que algumas vezes, inclusive, são apresentados como sinônimos, embora não sejam. São eles *crowdsourcing*, computação social e inteligência coletiva (Quinn and Bederson, 2011), que são definidos da seguinte forma:

- ***Crowdsourcing*** é o ato de terceirizar tarefas, tradicionalmente realizadas por um funcionário ou contratado da empresa, a um grupo grande, indefinido, de pessoas ou comunidade (*crowd*) através de um convite aberto (Howe, 2006).

- **Computação social** é um termo geral para uma área da ciência da computação que está preocupada com a intersecção de comportamento social e sistemas computacionais. A computação social utiliza a tecnologia para fins sociais facilitando a ação coletiva e interação social on-line. As pessoas podem gerar, distribuir e compartilhar informações para grupos de pessoas através de blogs, *wikis* e comunidades on-line, por exemplo. Dessa forma, a finalidade do ser humano nesse contexto não é atuar como um processador humano, mas interagir com outras pessoas ou contribuir com opiniões.

- **Inteligência coletiva** A inteligência coletiva é um conceito que descreve um tipo de inteligência compartilhada que surge da colaboração de muitos indivíduos considerando suas diversidades. É uma inteligência distribuída por toda parte, na qual todo o saber está na humanidade, já que, ninguém sabe tudo, mas todos sabem alguma coisa (LEVY, 1999). Logo, são grupos de pessoas que fazem coisas que parecem coletivamente inteligentes (Little et al., 2010).

A Figura 2.1 apresenta a relação entre esses termos. A inteligência coletiva é apresentada como um superconjunto de computação social, computação por humanos e *crowdsourcing*.

Figura 2.1: Diagrama de Venn para ilustrar como os termos computação por humanos, *crowdsourcing* e computação social se relacionam. Adaptado de Quinn e Bederson (2011) .

A relação entre a computação por humanos e *crowdsourcing* está no fato de que os seres humanos que substituem os computadores na computação por humanos são substituídos por uma quantidade indefinida e desconhecida de seres humanos em *crowdsourcing*. Logo, *crowdsourcing* pode ser entendido como um método ou ferramenta utilizado para distribuir tarefas, facilitando a computação por humanos.

A interseção de *crowdsourcing* com computação por humanos representa as aplicações que poderiam ser realizadas ou por um humano ou por um computador digital. Por exemplo, a tradução é uma tarefa que pode ser feita por computadores quando a velocidade e custo são a prioridade, ou por seres humanos com proficiência no idioma quando a qualidade é a prioridade. No entanto, nem sempre a computação por humanos precisa usar *crowdsourcing*.

Isso acontece, por exemplo, quando tarefas são distribuídas para um conjunto restrito de trabalhadores contratados através do processo de recrutamento tradicional.

A interseção da computação por humanos com computação social inclui as aplicações que necessitam da capacidade humana de grupos de pessoas que podem interagir entre si, isto é, da contribuição colaborativa de grupos de pessoas. As redes sociais podem ajudar em pesquisas de mercado e avaliação de produto, por exemplo.

As aplicações que são *crowdsourcing* e não são computação social são aquelas em que não há interação entre as pessoas. Por exemplo, ideias ou projetos não são necessariamente compartilhados com outros participantes. E as redes sociais que apenas facilitam a comunicação entre os seus membros sem terceirizar atividades para seus membros, não podem, portanto, ser classificadas como iniciativas de *crowdsourcing*.

É possível então perceber que os problemas tratados pela computação por humanos, no âmbito computacional, se encontram em aplicações *crowdsourcing* e computação social.

2.2 Sistemas de computação por humanos

Um sistema de computação por humanos pode ser visto como um sistema distribuído que agrega o poder computacional de um grupo de seres humanos conectados a Internet. Os seres humanos são denominados ou de **solicitantes** ou de **trabalhadores**, e são os dois principais usuários finais dos sistemas de computação por humanos.

O sistema de computação por humanos funciona, portanto, como uma interface entre os solicitantes e os trabalhadores, abstraindo, para ambos, a complexidade envolvida na gerência das tarefas disponibilizadas pelos solicitantes e na oferta da capacidade cognitiva dos trabalhadores.

Existem várias aplicações de computação por humanos distribuída, mas os mais comuns são os jogos com propósito (Ahn, 2005), sistemas de pensamento distribuído (Ponciano et al., 2014) e sistemas de trabalho on-line (Ipeirotis, 2010b).

A ideia dos jogos com propósito é que as pessoas contribuam com a solução de um problema enquanto se divertem jogando. Nesse caso, o trabalho do ser humano é implícito. O termo foi criado por Luis von Ahn [(Ahn, 2005) (Von Ahn and Dabbish, 2008). Um de seus jogos é o ESP que tem como objetivo rotular uma grande coleção de imagens com palavras-chave a fim de descobrir se os participantes podem ajudar a criar pesquisa de imagens mais precisas e acessibilidade para usuários com deficiência visual (Von Ahn and Dabbish, 2004). Consiste em solicitar a dois usuários aleatórios para, de forma colaborativa, rotular uma imagem, finalizando quando ambos os usuários fornecem um rótulo comum. Outros exemplos de jogos com propósito incluem o Peekaboom (Von Ahn et al., 2006), Verbosity (Von Ahn et al., 2006), Tag-A-Tune (Law et al., 2007) e Foldit (Cooper et al., 2010).

Outro projeto que usa o trabalho implícito é o reCAPTCHA que utiliza as pessoas para resolverem CAPTCHAs, imagens distorcidas de texto que são usadas em sites para impedir o acesso de programas automatizados (Von Ahn et al., 2008). Ao solucionar os CAPTCHAS as pessoas ajudam a transcrever livros antigos digitalizados e jornais que não podem ser processados pelo reconhecimento ótico de caracteres (OCR), devido ao envelhecimento. Milhões de CAPTCHAs são resolvidos diariamente.

Nos sistemas de pensamento distribuído e mercados de trabalho on-line o foco dos trabalhadores é executar tarefas de computação por humanos. Nesse caso, o trabalho do ser humano é explícito. Além disso, esses sistemas permitem uma maior diversidade de aplicações em relação aos anteriores.

Sistemas de pensamento distribuído permitem que as pessoas realizem as tarefas como um trabalho voluntário, sem receber nenhuma remuneração em troca, como por exemplo, em projetos de ciência cidadã. A motivação nesse caso está relacionada ao prazer de colaborar e de aprender. O exemplo mais conhecido é o Zooniverse[1], uma plataforma on-line de ciência cidadã onde os voluntários completam tarefas de pesquisa relacionadas a disciplinas como astronomia, biologia celular, ecologia e humanidades. Exemplos de projetos da plataforma Zooniverse incluem o Galaxy Zoo, cujo objetivo é classificar as galáxias de acordo com a forma (Lintott et al., 2008), e Planet Hunters, cujo objetivo é identificar planetas extrassolares

[1] http://www.zooniverse.org

a partir das curvas de luz das estrelas registradas pelo telescópio espacial Kepler (Fischer et al., 2012).

Nos mercados de trabalho on-line as pessoas realizam as tarefas em troca de uma compensação, financeira ou social (Yuen et al., 2009; Ipeirotis, 2010b; Yuen et al., 2011). Nesses mercados o problema do solicitante é dividido em tarefas que são distribuídas para os trabalhadores através da plataforma. Os resultados obtidos são agregados e entregues ao solicitante. Estes mercados de trabalho on-line podem ser classificados como mercados de propósito geral, como por exemplo, o Amazon Mechanical Turk, oDesk, Freelancer e Crowdflower, bem como mercados mais especializados como TopCoder, uTest e 99Designs, voltados para competição de programação, teste de software e concurso de design de novos produtos, respectivamente.

Os mercados de trabalho on-line e pensamento distribuído são mais genéricos do que os jogos com propósito porque suportam uma maior diversidade de aplicações enquanto que os jogos com propósito são desenvolvidos para uma aplicação específica consumindo, assim, mais tempo e recurso financeiro (Mao et al., 2011).

2.3 *Crowdsourcing* remunerado

O termo *crowdsourcing* surgiu em 2006 num artigo publicado por Jeff Howe na revista americana Wired para descrever um novo modelo de negócios (ou resolução de problemas) que aproveita soluções criativas de uma rede distribuída de pessoas conectadas a Internet e que colabora de forma voluntária (Howe, 2006). Resultante da contração das palavras *"crowd"* e *"outsourcing"*, o *crowdsourcing* significa, em essência, o ato de recorrer à multidão.

Quinn e Bederson de forma mais abrangente, definem *crowdsourcing* como *"o ato de explorar as habilidades perceptivas, inativas e cognitivas de muitas pessoas para alcançar um resultado bem definido, como resolver um problema, classificar um conjunto de dados, ou produzir uma decisão"* (Vakharia and Lease, 2013). Esta definição não relaciona diretamente *crowdsourcing* com empresa ou trabalhador, nem com tecnologia específica, mas com a

ideia do poder das multidões, assim como as definições encontradas nos trabalhos de Kittur, Chi e Suh (2008) e Doan, Ramakrishnan e Halevy (2011). É a computação por humanos distribuída.

Considerando que a computação por humanos substitui computadores por seres humanos, *crowdsourcing* substitui trabalhadores humanos tradicionais, contratados formalmente pelas empresas, por um grupo indefinido de pessoas desconhecidas, ou seja, a multidão, em função da demanda de trabalho. Além disso, em *crowdsourcing*, o trabalho a ser realizado é selecionado pelos trabalhadores e não pelo empregador, como ocorre no modelo tradicional de trabalho. O termo tem sido usado para uma grande variedade de fenômenos e está relacionado a áreas como inovação aberta, criação colaborativa ou geração de conteúdo (Schulze et al., 2011).

O *crowdsourcing* remunerado acontece quando os trabalhadores realizam o trabalho em troca de recompensa financeira, isto é, os solicitantes (indivíduos, grupos ou organizações) recompensam financeiramente os trabalhadores que realizam as tarefas em um mercado de trabalho on-line (Kittur et al., 2013). Empresas como a CrowdFlower [2] e CloudCrowd [3] deram visibilidade ao termo em função do montante de financiamento recebido[4,5]. Frei define *crowdsourcing* remunerado como o uso de uma tecnologia intermediária para a terceirização de trabalho remunerado, de todos os tipos, para um grande grupo de trabalhadores (Frei, 2009).

O trabalho submetido em plataformas *crowdsourcing*, considerando a granularidade, isto é, a quantidade e o tamanho das tarefas, é classificado em quatro categorias assim descritas [47]:

- Microtarefas: as principais características são volume muito alto de tarefas, recompensa muito baixa por tarefa e tarefas fortemente automatizadas. São exemplos de microtarefas: pesquisar endereço de e-mail ou sites de empresas, traduzir descrição de produto, pesquisar preços de produtos e categorizar produtos.

[2] www.crowdflower.com
[3] www.cloudcrowd.com
[4] http://techcrunch.com/2010/01/20/crowdflower-raises-5-million-for-cloud-sourced-labor/
[5] http://techcrunch.com/2010/08/13/cloudcrowd-raises-5-1-million-to-outsource- labor-to-the-cloud/

- Macrotarefas: geralmente possui alto volume de tarefas, taxa de remuneração baixa e processo automatizado. São exemplos de tarefas nessa categoria: escrever revisão de produto, testar site e prover feedback, completar citação de pesquisa, pesquisar instituições que fazem pesquisa em determinado tema dentre outras.

- Projetos simples: baixo volume ou tarefas individuais com moderada recompensa, em geral exigindo algum contato com o trabalhador. Criar um site de marca, preparar um esboço de apresentação para uma conferência e contactar participantes confirmados para um evento são exemplos de tarefas nessa categoria.

- Projetos complexos: projeto único, com alta taxa de remuneração, normalmente exigindo uma quantidade substancial de tempo e interação direta com o trabalhador. Programar um módulo de software e desenvolvimento de produtos são exemplos.

É possível perceber que as categorias variam em relação ao valor da recompensa, interação entre solicitante e trabalhador, isto é, desde a falta de interação até a interação direta, automatização do processo e dificuldade das tarefas.

A recompensa pode ser bem significativa quando *crowdsourcing* é realizado através de concurso/desafio porque apenas as melhores soluções são recompensadas. Por exemplo, a NASA[6] lançou uma série de concursos na plataforma TopCoder que oferecia US$35.000 em prêmios no período de seis meses aos cientistas que desenvolvessem os melhores algoritmos para identificar asteróides em imagens de telescópios.

A recompensa é pequena quando o solicitante tem um problema que pode ser particionado em várias pequenas tarefas, ou microtarefas, que podem ser realizadas em alguns minutos. Nesse caso, a recompensa não é oferecida apenas para as melhores soluções e sim para todas as soluções aprovadas pelo solicitante, permitindo assim, que o trabalhador possa aumentar sua recompensa financeira, dado que a recompensa é por microtarefa. São exemplos de plataformas de mercado de trabalho on-line de microtarefas o MTurk, microWorkers[7], oDesk[8]

[6]http://www.topcoder.com/asteroids/asteroiddatahunter/
[7]https://microworkers.com
[8]https://www.odesk.com

e clickworker[9], dentre outras.

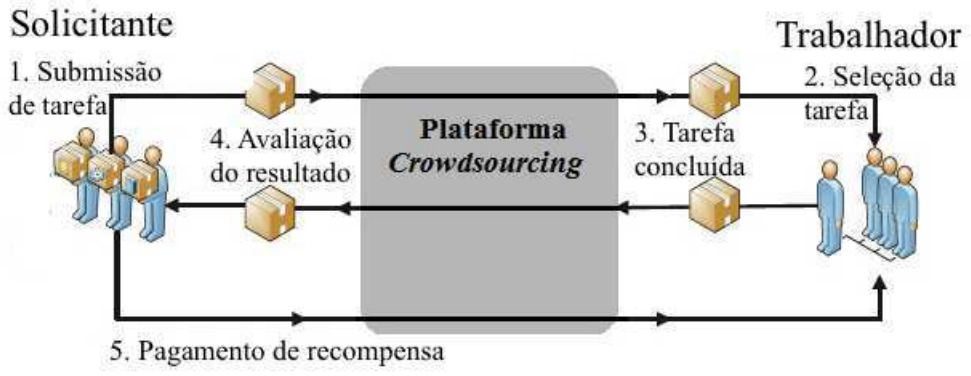


Figura 2.2: Esquema de funcionamento de plataforma *crowdsourcing*. Adaptado de Hirth et al.(2013).

A Figura 2.2 apresenta, de forma esquemática, o funcionamento de plataformas *crowdsourcing* remunerado, que inclui as seguintes atividades: (1) o solicitante envia uma tarefa para a plataforma *crowdsourcing*, definindo a recompensa que será paga e como o trabalho será avaliado; (2) o trabalhador escolhe as tarefas que deseja realizar; (3) as tarefas concluídas são enviadas para avaliação; (4) o solicitante avalia o resultado enviado; (5) se o trabalho for aceito, o solicitante remunera o trabalhador.

2.4 Trabalhadores, solicitantes e motivação

O sucesso de um projeto em computação por humanos depende do recrutamento dos trabalhadores e de sua permanência no mesmo. Por isso, é importante conhecer os motivos que impulsionam a participação dos solicitantes e trabalhadores, que podem ser diversos e distintos.

Estar motivado significa agir para atingir um ou mais objetivos. Uma pessoa que não sente nenhum impulso ou vontade para agir é, portanto, caracterizado como desmotivado,

[9]http://clickworker.com/en/

enquanto alguém que está energizado ou ativado em direção a um objetivo é considerado motivado (Ryan and Deci, 2000).

As pessoas que têm motivação não só apresentam diferentes quantidades de motivação como também diferentes tipos de motivação. A Teoria da Autodeterminação (Deci and Ryan, 2013) distingue dois tipos diferentes de motivação com base nas diferentes razões ou objetivos que dão origem a uma ação: a motivação intrínseca e a motivação extrínseca. A motivação intrínseca se refere a fazer algo porque é inerentemente interessante ou agradável, e a motivação extrínseca refere-se a fazer alguma coisa para obter algum resultado.

Convém ressaltar que é falso dizer que a motivação extrínseca é fruto da ação do ambiente e a intrínseca da pessoa, porque, a motivação é sempre fruto de uma interação entre a pessoa e o ambiente no qual está inserido. Segundo Ryan e Deci (2000) a motivação intrínseca é uma motivação inata porque as pessoas, desde o nascimento, são ativas e curiosas, apresentando disposição para aprender e explorar. Porém, à medida que a pessoa cresce, passa a conviver com regras sociais, a ter que realizar tarefas que nem sempre são prazerosas e interessantes, logo as pessoas não realizam as tarefas porque estão motivadas intrinsecamente.

Pesquisas mostram que a qualidade da experiência e desempenho pode ser muito diferente quando uma pessoa age movida pela motivação intrínseca e quando movida pela motivação extrínseca (Archak, 2010). No entanto, observa-se que os dois tipos de motivação podem aparecer mesclados, como, por exemplo, quando a pessoa estuda um tema que a interessa (motivação intrínseca) e consegue com isso um bom resultado (motivação extrínseca) (Heckhausen, 1991).

Quando as pessoas envolvidas em uma tarefa estão suficientemente motivadas, elas conseguem superar quaisquer tipos e graus de dificuldade (Bueno, 2002). A relação entre motivação e desempenho pode ser explicada pelo fato de que todas as ações precisam de uma motivação e quando elas estão diretamente ligadas às realizações pessoais, atendendo seus sonhos e expectativas, elas impulsionam ainda mais seu desempenho, produzindo mais e melhor.

Do ponto de vista do solicitante, a motivação primeira é a necessidade de resolver algum problema computacional com a maior precisão e eficiência, minimizando custos.

Do ponto de vista do trabalhador, em *crowdsourcing* remunerado, a principal motivação parece ser a recompensa financeira, isto é, a motivação extrínseca. No entanto, outros fatores podem influenciar a decisão dos trabalhadores e influenciar de forma diferente os vários tipos de trabalhadores (Brabham, 2008; Lakhani et al., 2007; Kaufmann et al., 2011; Rogstadius et al., 2011). O trabalhador pode desejar executar uma tarefa em função de motivação intrínseca, isto é, pelo prazer que a tarefa lhe proporciona ou pelo aprendizado que pode obter ao executá-la ou pela oportunidade de contribuir para o bem comum ou aprender algo novo.

Howe (2006) afirma que "ao contrário do que diz a mentalidade convencional, o ser humano nem sempre se comporta seguindo padrões egoístas". As pessoas são capazes de colaborar por pouca ou nenhuma remuneração, motivadas pelo desejo de beneficiar uma comunidade, de fazer um bem maior, pelo prazer de praticar seu oficio, de se superar, pelo prazer em cultivar os próprios talentos e partilhar o que conhecem. Assim, a colaboração é a própria recompensa.

Portanto, conhecer os motivos que levam os trabalhadores a participarem de atividades *crowdsourcing* fornece aos solicitantes mais subsídios para escolher a melhor estratégia para o recrutamento de trabalhadores e qual motivação (ou combinação de motivações) deve ser mais incentivada na tarefa para obtenção de melhores resultados.

2.5 Considerações finais

Este capítulo forneceu uma visão geral sobre o paradigma computação por humanos e termos relacionados encontrados na literatura, além de apresentar os sistemas de computação por humanos, os participantes (trabalhadores e solicitantes) e os fatores motivacionais que podem impulsionar a atividade dos participantes em tais sistemas.

O termo *crowdsourcing* se destaca porque, além de facilitar a computação por humanos, introduz uma nova forma de relação de trabalho que é a possibilidade de utilizar uma mul-

tidão de pessoas, desconhecidas, e que geograficamente podem estar bem dispersas, para solucionar um problema a um custo pequeno.

Apesar do número de participantes em mercados de trabalho on-line ser crescente, utilizar a multidão como força de trabalho, no entanto, também tem seus problemas. Nessas plataformas, os trabalhadores não precisam ter habilidade específica para participar e, além disso, muitos têm como principal objetivo aumentar o rendimento através das recompensas oferecidas. Como consequência, o solicitante recebe resultados de baixa qualidade. Obviamente que alguns resultados ruins são fruto de outros fatores como a ignorância do trabalhador, isto é, a falta de conhecimento suficiente para realizar determinada tarefa, ou mesmo, por um deslize como, por exemplo, falta de entendimento correto sobre o que deve ser feito. Para conseguir melhores resultados, o solicitante deve utilizar mecanismos para prevenir tais trabalhadores.

Neste trabalho, dá-se ênfase ao uso de testes de qualificação para pré-selecionar trabalhadores em mercados de trabalho on-line, de microtarefas. O estudo de mecanismos de controle de qualidade é apresentado no próximo capitulo, seguido dos estudos realizados em plataforma *crowdsourcing* selecionada.

Capítulo 3

A qualidade dos resultados em mercados de trabalho on-line e a plataforma MTurk

O sucesso de um projeto em computação por humanos depende, em primeiro lugar, do recrutamento dos trabalhadores e de sua permanência no mesmo. Em segundo lugar, que as respostas obtidas sejam confiáveis.

Em mercados de trabalho on-line de microtarefas, ao contrário de plataformas mais especializadas, os trabalhadores não precisam ter habilidade específica para participar. Além disso, muitos têm como principal objetivo aumentar o rendimento através das recompensas oferecidas. Quanto menor o esforço desprendido para executar uma tarefa, mais rapidamente a tarefa poderá ser concluída e maior será a recompensa obtida durante uma determinada sessão de trabalho. Por outro lado, a redução do esforço empregado na execução da tarefa pode comprometer a qualidade do resultado. Mesmo motivados, os trabalhadores podem tentar trapacear o sistema ou até mesmo resolver de forma incorreta a tarefa, por não ser capaz ou não entender o problema.

Algumas situações observadas são:

- Alguns trabalhadores apresentam resultados incorretos, através de respostas aleatórias, de forma a maximizar os seus rendimentos realizando o maior número de tarefas possí-

vel no menor tempo possível. Isso acontece porque os trabalhadores são remunerados, em geral com valor pequeno, por tarefa (Demartini et al., 2012). Logo, quanto mais tarefas realizadas, maior será a recompensa total. Como consequência, os resultados podem não ser os esperados. A verificação da corretude das respostas não é uma atividade fácil uma vez que os solicitantes utilizam este tipo de mercado com o intuito de obter resultados de muitos trabalhadores. Além disso, a verificação do resultado deve ser automatizada, o que não é fácil para tarefas subjetivas.

- A possibilidade de conluio, isto é, os trabalhadores combinam as respostas para as tarefas. Apesar de alguns sistemas *crowdsourcing* não permitirem interação entre os trabalhadores eles podem usar outros canais de comunicação, como os fóruns[1] [2], para combinar as respostas (Barowy et al., 2012). Nesse caso, o solicitante pode ser induzido a um erro se utilizar, por exemplo, a estratégia de votação para decidir qual a resposta certa. Os fóruns também são utilizados para combinar a não realização de tarefas de um dado solicitante, ou de forma positiva, para pesquisar informações sobre a especificação de tarefas (Kulkarni et al., 2012).

- Outra situação que possibilita resultados ruins são as tarefas mal projetadas ou confusas, isto é, tarefas que não possuem instruções claras o suficiente e que podem provocar ambiguidade no entendimento ou falta de entendimento do que deve ser feito [8], além de que podem gerar fadiga no trabalhador, aumentando a probabilidade de erros e reduzindo assim a produtividade. Esse caso, em geral, é provocado por solicitantes com pouca experiência no projeto de tarefas e que afeta também a qualidade dos trabalhadores bem intencionados.

É claro que uma pequena quantidade de resultados incorretos pode ser tolerada, como em qualquer tipo de sistema. No entanto, é evidente que o controle de qualidade é um desafio em *crowdsourcing* remunerado e é importante que exista para garantir que o problema do solicitante seja resolvido de forma aceitável.

[1] http://www.turkernation.com

[2] http://mturkforum.com/

No contexto de mercados de trabalho on-line de microtarefas ou *crowdsourcing* remunerado há uma série de pesquisas com foco no controle da qualidade do trabalho realizado pelos trabalhadores. A motivação para essas pesquisas é, portanto, a variância na qualidade dos resultados submetidos pelo trabalhador. As estratégias investigadas, em geral, dividem-se, basicamente, em dois campos complementares: as que analisam os resultados submetidos (Kittur et al., 2013) e as que têm o objetivo de prevenir resultados de baixa qualidade. Estas estratégias são apresentadas neste capítulo assim como possíveis lacunas que ainda podem ser exploradas no contexto.

Neste capítulo também é apresentada a plataforma Mechanical Turk (MTurk), especificamente o seu funcionamento e os detalhes dos mecanismos existentes que permitem aos solicitantes fazer uma pré-seleção dos trabalhadores. Na plataforma MTurk, os trabalhadores realizam pequenas tarefas em troca de remuneração. A plataforma MTurk foi escolhida em função de sua popularidade e tamanho. A plataforma MTurk foi liberada em 2005 e em 2010 já haviam 400 mil trabalhadores registrados (Ross et al., 2010). Os dados de agosto de 2014 indicam que a força de trabalho do MTurk consistia em mais de 500 mil trabalhadores espalhados em 190 países ao redor do mundo (Amazon Web Services, 2017).

3.1 Estratégias de controle de qualidade dos resultados

3.1.1 Controle de qualidade após submissão de resultados

O trabalhador pode ter seus resultados analisados somente após a submissão dos mesmos. Duas estratégias podem ser usadas nesse caso: uso de conjunto de teste (Kittur et al., 2008; Oleson et al., 2011; Le et al., 2010) ou esquemas de replicação (Barowy et al., 2012; Le et al., 2010; Callison-Burch, 2009; Snow et al., 2008).

A estratégia de conjunto de teste, ou *gold standard data*, consiste em comparar as respostas do trabalhador com as respostas corretas do conjunto de teste (Kittur et al., 2008). O conjunto de teste contém questões que são representativas da tarefa real e que são inseridas no conjunto de questões da tarefa. Se as respostas submetidas pelo trabalhador são significativamente diferentes das do conjunto de teste, o solicitante pode rejeitar automaticamente

as submissões do trabalhador ou treinar o trabalhador na tarefa. Assim, os solicitantes podem inferir automaticamente a acurácia dos trabalhadores computando a sua taxa de sucesso quando os resultados são apresentados.

Conjuntos de teste são fáceis de entender, explicar e programar, o que explica sua popularidade. São considerados uma armadilha poderosa especialmente quando as questões de teste não podem ser diferenciadas das questões regulares (Difallah et al., 2012). Alguns problemas relacionados a esta estratégia são:

- O custo. Para grandes quantidades de tarefas, é necessário um número maior de questões de teste.
- As questões de teste devem ser escolhidas com cuidado para não enganar os trabalhadores honestos e não serem fáceis para programas maliciosos (*bots*).
- O trabalhador pode se qualificar com poucas questões visto que o conjunto de teste é um subconjunto pequeno do total de questões. Logo, o método não identifica todos os trabalhadores maliciosos (Zhu and Carterette, 2010) como também pode prejudicar trabalhadores bem intencionados que porventura errem por outros fatores.
- Trabalhadores maliciosos podem mudar de comportamento sempre que desconfiarem estar diante de questões do conjunto de teste.
- Para tarefas subjetivas ou de geração de conteúdo não é viável usar essa estratégia.

A estratégia de esquemas de replicação para analisar os resultados dos trabalhadores é investigada em diversos estudos (Sheng et al., 2008; Barowy et al., 2012; Le et al., 2010; Callison-Burch, 2009; Snow et al., 2008; Alonso and Mizzaro, 2009). A replicação usa respostas para a mesma tarefa, submetidas por vários trabalhadores diferentes, para identificar e eliminar respostas incorretas. Usando regras como votação majoritária é possível identificar as respostas corretas com alta probabilidade.

A votação majoritária é utilizada geralmente para avaliar os resultados submetidos para tarefas de categorização de dados ou de anotação de imagem (Sorokin and Forsyth, 2008). Considera como correta a resposta de consenso da maioria dos diferentes trabalhadores que

submeteram resultados. O esquema de funcionamento é apresentado na Figura 3.1 conforme o seguinte fluxo: (1) o solicitante envia a tarefa para a plataforma; (2) a tarefa é replicada para vários trabalhadores; (3) trabalhadores diferentes realizam a mesma tarefa; (4) algoritmo de votação majoritária é executado; (5) resultado é devolvido para o solicitante. Cada trabalhador que vota de forma idêntica ao resultado de consenso recebe a remuneração.

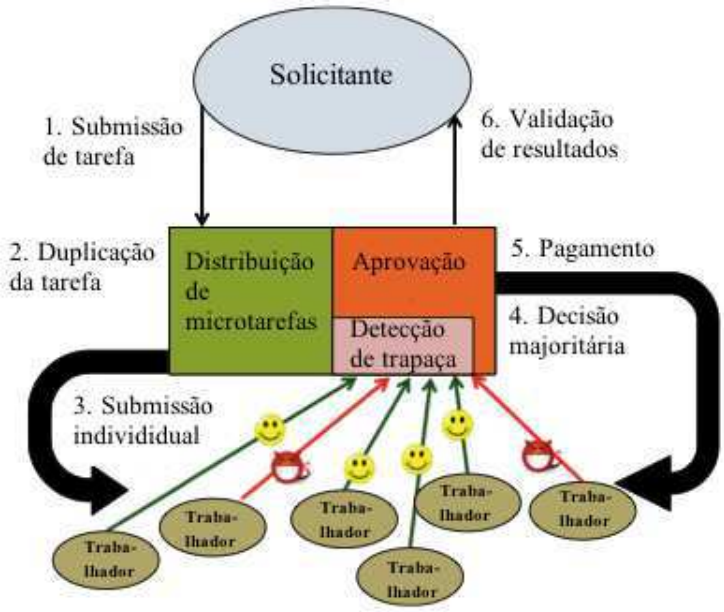


Figura 3.1: Esquema de funcionamento da votação majoritária. Adaptado de Hirth et al. (2013).

Embora possa apresentar bons resultados, como mostrado em trabalhos que apresentam experimentos realizados usando julgamento de relevância e anotação (Sheng et al., 2008; Snow et al., 2008; Alonso and Mizzaro, 2009), a votação majoritária tem seus problemas. Um possível ponto fraco é a suposição implícita de que todos os trabalhadores são igualmente bons (Snow et al., 2008), quando na verdade não são. Por exemplo, supondo a situação em que apenas um trabalhador é especialista na tarefa e, portanto, a probabilidade de responder corretamente é alta e, os demais trabalhadores não são e respondem de forma idêntica mas contrária ao trabalhador especialista, a votação majoritária favorece o erro.

Diversos estudos tem pesquisado alternativas para determinar o consenso da maioria (Khanna et al., 2010; Hirth et al., 2013; Callison-Burch, 2009; Raykar et al., 2010). Hirth

et al., por exemplo, analisam a votação majoritária utilizando um grupo de trabalhadores, denominado grupo de controle, que tem a função de validar a resposta submetida por um trabalhador para a tarefa (Hirth et al., 2013). É a chamada revisão por pares. A Figura 3.2 apresenta o funcionamento do esquema de revisão por pares, que funciona da seguinte forma: (1) o solicitante envia a tarefa para a plataforma *crowdsourcing*; (2) um trabalhador escolhe a tarefa e, (3) depois submete a resposta; (4) a plataforma gera novas tarefas de validação para o resultado submetido pelo trabalhador que é passado para o grupo de controle que utiliza critérios estabelecidos para julgar; (5) as classificações dos diferentes trabalhadores são devolvidos para a plataforma, que (6) calcula a classificação geral do resultado do trabalhador. O resultado do trabalhador será considerado válido se a maioria do grupo de controle assim decidir. Em caso positivo, (7) o trabalhador é remunerado; (8) o solicitante recebe o resultado. Em caso negativo, a tarefa é repetida por outros trabalhadores até que se obtenha um consenso. Nessa abordagem, as tarefas possuem remuneração diferente, sendo a tarefa principal melhor remunerada.

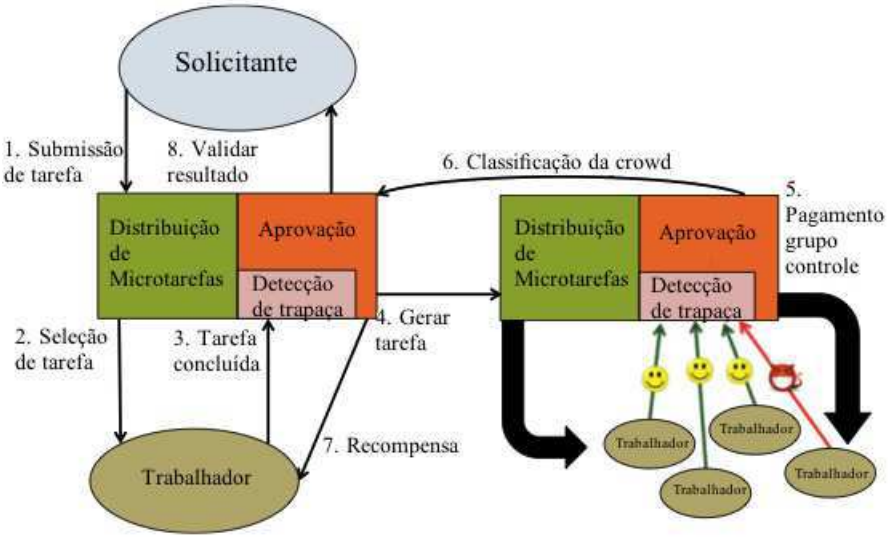


Figura 3.2: Esquema de funcionamento da estratégia de votação majoritária com grupo de controle. Adaptado de Hirth et al. (2013).

Hirth et al. (2013) avaliam que para tarefas mais simples a votação majoritária simples é mais conveniente e para tarefas subjetivas, como julgamento de relevância de artigos, é melhor usar o grupo de controle. Mesmo assim, ainda persiste o problema que essa estratégia

é suscetível ao conluio de trabalhadores (Kittur et al., 2013).

Nas duas abordagens apresentadas para avaliar a qualidade do trabalhador após a submissão dos resultados há um aumento no custo associado com a execução de uma tarefa, uma vez que se faz necessário ou aumentar o número de tarefas, para incluir o conjunto de teste, ou aumentar o número de trabalhadores, a fim de atingir o nível exigido de replicação. Assim, um equilíbrio deve ser estabelecido entre o custo extra induzido pela abordagem e sua eficácia para detectar automaticamente os resultados de baixa qualidade. Além disso, essas abordagens normalmente só se aplicam a tarefas cujas respostas podem ser consideradas como corretas ou incorretas, ou seja, a tarefas com um conjunto fechado de possíveis respostas. Elas não se aplicam a tarefas que envolvem a subjetividade ou criatividade (Dow et al., 2011).

3.1.2 Controle de qualidade preventivo

O controle da qualidade em plataformas *crowdsourcing* pode ser realizado através da prevenção, isto é, evitando que trabalhadores maliciosos submetam resultados de baixa qualidade. A ideia é projetar tarefas de tal forma que elas se tornem resistentes ao trabalho de baixa qualidade, tendo em conta várias características, tais como a definição de tarefas e interface de usuário (Kittur et al., 2008; Khanna et al., 2010; Ipeirotis et al., 2010; Amazon Web Services, 2017; Callison-Burch, 2009), o tempo de conclusão (Sorokin and Forsyth, 2008), e manipulação de incentivos (Rogstadius et al., 2011; Mason and Watts, 2010; Shaw et al., 2011). Os estudos mostram que estes aspectos afetam a qualidade dos resultados obtidos.

Eickhoff e De Vries, por exemplo, analisaram os métodos de votação majoritária e uso de conjunto de teste, considerando se o tipo da tarefa, o projeto da interface e seleção de trabalhadores, com base na localização geográfica, afetam o número de trabalhadores maliciosos (Eickhoff and de Vries, 2011). Com base nos resultados experimentais, concluíram que os trabalhadores maliciosos são menos frequentes em tarefas não repetitivas que envolvem um grau de criatividade e abstração, e que a seleção prévia dos trabalhadores pode reduzir muito o número de trabalhadores maliciosos.

A recompensa oferecida pelos solicitantes na maioria das tarefas é insignificante, mas mesmo assim percebe-se que os trabalhadores as executam. Por esse motivo, o incentivo financeiro é o método mais utilizado na perspectiva de obter melhor qualidade de trabalho. No entanto, a oferta de uma maior recompensa pode atrair mais trabalhadores e pode afetar a rapidez na conclusão da tarefa, mas não necessariamente aumentar a qualidade do resultado (Kittur et al., 2008). No Mechanical Turk, por exemplo, o valor oferecido para os trabalhadores realizarem as tarefas é muito pequeno o que resulta em salários baixos também (Ipeirotis, 2010c). Porém, aumentar o valor pode não ter o efeito positivo esperado pelo solicitante (Mason and Watts, 2010).

Uma alternativa é investir no projeto da tarefa. Dependendo do tipo e do contexto da tarefa, a apresentação pode influenciar a qualidade dos resultados de duas maneiras. Primeiro, um projeto bom e apropriado facilita a compreensão global do trabalho e, portanto, aumenta a chance de resultados corretos (Khanna et al., 2010). Segundo, certas propriedades do projeto podem influenciar não só a atratividade da tarefa, mas também o interesse de grupos específicos de trabalhadores. Schulze et al. (2011) realizaram estudos com os trabalhadores com o intuito de identificar algumas propriedades potencialmente relevantes concluindo que algumas delas dependem inclusive da origem do trabalhador. Por exemplo, trabalhadores americanos preferem tarefas de solicitantes que possuem boa reputação enquanto trabalhadores indianos preferem tarefas que pagam bônus.

Resultados de baixa qualidade, portanto, podem ser evitados se a tarefa é definida de forma adequada (Ipeirotis, 2010c). Seria a prevenção através de filtragem implícita. Sob esse ponto de vista, se os trabalhadores acham a tarefa mais envolvente, interessante, provavelmente os resultados seriam de melhor qualidade. Os trabalhadores estariam motivados intrinsecamente. Assim, formular as questões e instruções corretamente, assim como as restrições, são aspectos importantes (Corney et al., 2010). Outros aspectos também merecem destaque como a forma de incentivo (Shaw et al., 2011; DiPalantino and Vojnovic, 2009), o tempo para realizar a tarefa e as características dos trabalhadores (Hirth et al., 2011; Singh et al., 2002).

Uma interface de usuário complicada pode desencorajar os trabalhadores honestos e pode levar a atrasos (Kittur et al., 2008; Khanna et al., 2010; Ipeirotis et al., 2010; Allahbakhsh et al., 2013). Quando o tempo necessário para completar a tarefa não é apropriado, os trabalhadores podem se sentir pressionados e, assim, responder de qualquer forma, ou mudar para outras tarefas, diminuindo assim a sua atenção; em todos os casos, o resultado é uma redução na qualidade dos resultados produzidos (Eickhoff and de Vries, 2011). A principal questão em relação ao projeto da tarefa é que usada de forma isolada não garante a qualidade dos resultados e a falta de, ou pouca, experiência do solicitante leva a tarefas mal planejadas. Além disso, nem sempre as tarefas mais interessantes, que motivam mais os trabalhadores, garantem resultados de melhor qualidade (Khazankin et al., 2011). Portanto, projeto de tarefa e manipulação de incentivos formam uma boa barreira para trabalhadores maliciosos, mas constituem um fardo para os solicitantes.

Outra possibilidade é selecionar os trabalhadores para executar a tarefa, usando a reputação do trabalhador, ou seja, o histórico do seu comportamento na execução de tarefas anteriores. A reputação é construída, principalmente, através do feedback dos solicitantes sobre os trabalhadores ao sistema. Reputação pode ser usada para impedir que um trabalhador realize tarefas (Khazankin et al., 2011), ou para prever a qualidade dos resultados de um trabalhador (Ipeirotis et al., 2010; Rzeszotarski and Kittur, 2011; Kokkodis and Ipeirotis, 2013) ou a probabilidade de que o trabalhador completará a tarefa (Khanna et al., 2010). Ipeirotis et al., por exemplo, propõem acompanhar a qualidade de um trabalhador, assim o solicitante pode conhecer previamente a qualidade esperada dos resultados (Ipeirotis et al., 2010). O método consiste em separar erros sistêmicos de viés devido a, por exemplo, uma distração, que pode prejudicar a reputação do trabalhador.

Apesar de ser um método bem utilizado pelos solicitantes não é uma estratégia adequada porque possui dois potenciais problemas que devem ser considerados (Eickhoff and de Vries, 2011). Esses problemas estão relacionados ao fato de que a reputação é calculada, principalmente, através do percentual de trabalho aceito e o número de tarefas executadas. O primeiro problema está relacionado ao fato de que os solicitantes aceitam todas as respostas e só as filtram posteriormente. Nesse caso, os trabalhadores maliciosos já receberam atualização de

precisão e mais tarefas poderão executar. O segundo problema ocorre quando o trabalhador utiliza o artifício de se cadastrar na plataforma *crowdsourcing* como solicitante, criar uma tarefa com muitas subtarefas e realizá-las imediatamente (Ipeirotis, 2010a). Apesar de ter um custo pequeno associado (percentual da comissão cobrada pela plataforma *crowdsourcing*), dessa forma o trabalhador aumenta sua pontuação na plataforma. Logo, trabalhadores com alta pontuação em trabalhos aceitos podem ser trabalhadores maliciosos.

Além disso, os sistemas de reputação em plataformas *crowdsourcing* apenas fornecem informações sobre o desempenho do trabalhador em relação aos trabalhos anteriores considerando o percentual de trabalho submetido, aceito, rejeitado ou abandonado, e não possui informação sobre as habilidades dos trabalhadores para realizar determinados tipos de tarefas (Amazon Web Services, 2017). Por exemplo, um trabalhador que tem experiência com anotação de imagem não necessariamente tem qualificação para trabalhar com tradução de texto também e, mesmo assim a qualidade de seu trabalho pode oscilar no decorrer do tempo (Schulze et al., 2013).

A habilidade dos trabalhadores pode ser avaliada através de mecanismo chamado de teste de qualificação, existente em algumas plataformas de *crowdsourcing* como CrowdFlower, oDesk e Mechanical Turk. O teste de qualificação permite que os solicitantes direcionem suas tarefas para grupos específicos de trabalhadores.

Vakharia e Lease (2013) compararam o mecanismo de teste de qualificação nos mercados mais importantes que fornecem esse recurso. O sistema CrowdFlower testa trabalhadores principalmente via conjunto de teste de questões. No entanto, os solicitantes também podem aplicar restrições de habilidades em tarefas por meio de testes de habilidade padronizados fornecidos pela plataforma. CrowdFlower usa *badges* para mostrar as habilidades dos trabalhadores, mas o interesse e a experiência do trabalhador não é registrado no sistema. A plataforma oDesk segue o modelo de contratação tradicional, permitindo que os solicitantes selecionem os trabalhadores por meio de entrevistas virtuais. Testes podem ser realizados pelos trabalhadores para construir a credibilidade, e os perfis dos trabalhadores incluem auto relato da proficiência em inglês e as áreas de interesse. No Mechanical Turk não há testes comuns pré-definidos para verificar habilidades ou para medir a proficiência em idioma, mas

a plataforma permite que os solicitantes desenvolvam seus próprios testes de qualificação (Amazon Web Services, 2017). Solicitantes também podem contar com o sistema de reputação do sistema ou utilizar trabalhadores pré-qualificados denominados de *mestres*, apesar de não ter conhecimento sobre quem são esses trabalhadores e nem como essa qualidade foi aferida. Embora não exista uma biblioteca com qualificações, um solicitante pode usar uma qualificação existente que foi criada por outro solicitante.

Apesar dos testes de qualificação permitirem ao solicitante pré-selecionar os trabalhadores em função dos requisitos de suas tarefas, também podem desencorajar trabalhadores com potencial que não estão dispostos a investir tempo em um teste com perspectivas desconhecidas de recompensa futura (Wais et al., 2010). Segundo Wais et al. (2010), o baixo percentual de trabalhadores recrutados pelo teste de qualificação utilizado em seus experimentos (apenas 35,6%) deve-se ao fato de que muitos trabalhadores tentam obter acesso rápido às tarefas e não se esforçam no teste, mesmo quando o teste não é difícil e possui questões de fácil compreensão. Também existe a possibilidade do trabalhador se esforçar no teste e depois reduzir seu esforço na tarefa real (Schulze et al., 2013).

No entanto, o teste de qualificação é um mecanismo conhecido e bem considerado pelos trabalhadores em relação aos demais mecanismos que avaliam os resultados após a submissão (Schulze et al., 2013) e há indícios de que podem melhorar de forma considerável a qualidade dos resultados (Su et al., 2007). Dessa forma, a estratégia de usar teste de qualificação pelos solicitantes parece ser o caminho mais adequado visto que não há receio de rejeição por parte dos trabalhadores.

3.2 Mechanical Turk

O Mechanical Turk (MTurk) é uma plataforma de distribuição de serviços da empresa Amazon que permite conectar solicitantes e trabalhadores através de uma interface gráfica na web (Kittur et al., 2008). É o exemplo mais importante de mercado *crowdsourcing* de microtarefas. Trabalhadores são indivíduos, anônimos, que aceitam realizar tarefas em troca de recompensa financeira. Solicitantes representam os indivíduos ou grupo de indivíduos

ou empresas que submetem pequenas tarefas na plataforma que podem ser concluídas em poucos minutos pelos trabalhadores.

Diversas pesquisas foram realizadas com o objetivo de analisar o aspecto demográfico dos trabalhadores da plataforma MTurk. Os resultados mostraram que a maioria dos trabalhadores eram basicamente americanos (47%) e indianos (34%), e que 90% das tarefas tinham uma recompensa de menos de 10 centavos de dólares, sendo o valor mais ofertado US$0,01 (Ipeirotis, 2010c,b). Difallah (2015) mostra que, com o passar do tempo, o número de trabalhadores indianos diminuiu enquanto que o número de trabalhadores canadenses aumentou, apesar de americanos e indianos ainda predominarem e, que a recompensa mais ofertada tem um valor de US$0,05.

As tarefas mais comuns no MTurk são transcrição e tradução de texto e áudio, coleta de opiniões, pesquisa de informação na web, classificação de imagens, música e documentos e, análise de sentimentos, por exemplo. Outras possibilidades são exploradas em alguns trabalhos, como por exemplo, Little et al. (2010) que apresentam diversos exemplos de uso que incluem escrita interativa, reconhecimento de texto ótico, experimentos relacionados à teoria de decisões por voto e experimentos psicológicos de reação a estímulos. Berstein (2015) apresenta o sistema Soylent que propõe o uso do MTurk para ser uma interface de edição de texto que permite trabalhadores do MTurk revisar e editar trechos de texto sob demanda.

As principais vantagens oferecidas pelo MTurk (Amazon Web Services, 2017) são:

- Força de trabalho sob demanda: centenas de milhares de trabalhadores em quase duas centenas de países disponíveis.
- Força de trabalho escalável: sem tamanho mínimo de projeto; é possível ter 100 tarefas em um dia e 10 mil no dia seguinte.
- Rapidez: trabalhadores podem trabalhar em paralelo.

3.2.1 Funcionamento do MTurk

O sistema *crowdsourcing* no MTurk funciona de forma idêntica ao apresentado na Figura 2.2 (seção 2.3), ou seja: o solicitante formula a tarefa a ser realizada pelos trabalhadores, cada trabalhador escolhe a tarefa que deseja realizar e, se a tarefa for concluída com êxito e aprovada pelo solicitante, o trabalhador recebe a recompensa. Nesse modelo de sistema, tanto os trabalhadores como os solicitantes são anônimos, portanto, não há interação direta entre eles e, nem entre os trabalhadores. Uma possibilidade de comunicação entre os trabalhadores pode ser através de fóruns.

O solicitante pode projetar as tarefas ou utilizando os recursos da interface gráfica, que também permite publicar e gerenciar tanto as tarefas como os trabalhadores, ou através da interface de linha de comando[3] ou usando a API[4] (*Application Programming Interface*) para automatizar a interação com o sistema (Amazon Web Services, 2017).

As tarefas criadas pelos solicitantes são chamadas de HITs (*Human Intelligence Task)* e são submetidas, em geral, em forma de grupo de HITs (*job*). A quantidade de HITs disponíveis no MTurk é da ordem de 100 mil [5]. Um grupo de HITs, portanto, pode conter 1 ou mais HITs, que exige o mesmo tipo de trabalho, no entanto, com dados diferentes. Por exemplo, a tarefa de descrever fotografias é a mesma, mas as fotos que devem ser descritas são diferentes em cada HIT.

O grupo de HITs possui, além dos HITs, um título, palavras-chave, uma descrição, a data de expiração, uma recompensa, o tempo necessário para a execução de cada HIT e a quantidade de HITs disponíveis no grupo de HITs. Opcionalmente, os solicitantes podem pré-selecionar trabalhadores para suas tarefas, associando às tarefas as qualificações que esses trabalhadores precisam cumprir - em alguns casos, os trabalhadores não podem sequer inspecionar os HITs de uma tarefa, se eles não possuem as qualificações exigidas. Nesse caso, as qualificações exigidas pelo solicitante também são parte integrante do grupo de HITs.

[3] http://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkCLT/Welcome.html
[4] http://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkAPI/Welcome.html
[5] http://www.mturk-tracker.com/#/general, último acesso em 26 de maio de 2016.

Um HIT deve ser simples, com objetivo bem definido e de fácil execução. Cada solicitante é responsável pelos HITs que cria, isto é, desde o projeto de interface do seu HIT, valor da recompensa oferecida, tempo necessário para a realização do mesmo até a avaliação das respostas submetidas pelos trabalhadores.

O MTurk disponibiliza para o trabalhador um quadro de tarefas (Figura 3.3) que permite que o trabalhador pesquise HITs por data de criação, quantidade de HITs disponíveis no grupo de HITs, valor da recompensa, data de expiração, título, tempo estimado para a execução, nome do solicitante ou qualquer outra palavra-chave. Este quadro de tarefas chama-se *HITs job board*.

Figura 3.3: Quadro de tarefas do MTurk utilizado pelos trabalhadores para seleção de tarefas. Fonte: www.mturk.com

Pesquisar HITs no quadro de tarefas do MTurk não é uma atividade fácil dado que as informações são atualizadas constantemente, isto é, o número de HITs disponíveis no MTurk muda a cada instante, podendo aumentar ou diminuir. Isso acontece, é claro, porque enquanto algumas tarefas estão sendo finalizadas, outras estão chegando no sistema. Além disso, a forma padrão de apresentar os HITs no MTurk é colocar os mais novos na primeira página do quadro de tarefas, deslocando assim, os HITs mais antigos para as páginas seguintes. No entanto, os HITs podem ser classificados, em ordem crescente ou decrescente, em seis categorias: pela data de criação (mais novo ou mais velho), pela quantidade dispo-

nível (mais ou menor número), pelo valor da remuneração (maior ou menor), pela data de vencimento (mais rápido ou mais recente), pelo título (a-z ou z-a), pelo tempo de execução (mais curto ou mais longo). Essas informações são disponibilizadas para cada grupo de HITs (Figura 3.3).

Quando o trabalhador encontra uma tarefa do seu interesse, ele tem a flexibilidade de aceitar ou não um HIT, como também aceitar e abandonar a qualquer momento. Quando o trabalhador aceita um HIT, este é associado exclusivamente ao trabalhador, isto é, o HIT se torna indisponível para outros trabalhadores. Ao finalizar a execução do HIT, o trabalhador submete o HIT para a apreciação do solicitante. Quando o HIT é aprovado pelo solicitante, o trabalhador recebe a recompensa financeira oferecida. No caso do HIT ser rejeitado pelo solicitante, o trabalhador não recebe a recompensa financeira. Os solicitantes podem oferecer bônus para os trabalhadores que submetem trabalhos considerados de alta qualidade. Se o grupo de HITs ainda possui HITs disponíveis, novos HITs são oferecidos ao trabalhador, que se aceitar, continuará a executar HITs de mesmo tipo.

A flexibilidade dada ao trabalhador em relação ao abandono de HIT, assim como quando ele aceita um HIT e não executa de forma satisfatória, tem seu ônus. O MTurk possui um sistema de reputação que atribui pontuação para os trabalhadores com base em suas respectivas estatísticas, representando assim, a experiência do trabalhador com o sistema. As estatísticas são ajustadas em função da participação do trabalhador no sistema, isto é, quantidade de HITs submetido, aprovado, rejeitado e abandonado, afetando a reputação do trabalhador, positivamente ou não.

3.2.2 Testes de Qualificação no MTurk

O MTurk oferece ao solicitante a possibilidade de selecionar trabalhadores para seus HITs através de testes de qualificação. Teste de qualificação, ou simplesmente qualificação, é uma propriedade que é concedida ao trabalhador para representar a sua habilidade, condição ou reputação. Pode ser um valor opcional, como um número. Assim o solicitante pode escolher quem pode e quem não pode participar de seus HITs através das qualificações. O solicitante pode, inclusive, utilizar mais de uma qualificação em seus HITs. Nesse caso, os

trabalhadores tem que atender a todas às qualificações e apenas os trabalhadores que possuem as qualificações exigidas pelo solicitante podem realizar os HITs.

O MTurk fornece testes de qualificação integrados ao sistema que usam as medidas de reputação do sistema ou informações pessoais do trabalhador e que estão disponíveis para todos os solicitantes.

O solicitante pode também criar novos testes de qualificação. Os testes de qualificação criados pelos solicitantes possibilitam aos mesmos um maior controle e flexibilidade na avaliação das habilidades dos trabalhadores para realização de seus HITs. A plataforma permite ao solicitante decidir se a qualificação deve ou não ser concedida ao trabalhador que realizou o teste de qualificação, determinando uma pontuação para o mesmo. Isso pode ser feito manualmente ou de forma automática. Além disso, a qualificação também pode ser concedida sem a necessidade de um teste associado, isto é, com base apenas na solicitação feita pelo trabalhador. Nesse caso, uma pontuação padrão é associada ao trabalhador que solicitou a qualificação. O solicitante que criou a qualificação pode posteriormente alterar essa pontuação, com base no desempenho futuro do trabalhador. Finalmente, os pedidos de qualificação para um tipo específico podem ser concedidos pelo solicitante a um trabalhador em particular, mesmo que esse trabalhador não tenha solicitado a qualificação. Basta que o solicitante conheça a identificação do trabalhador na plataforma.

Apesar do MTurk não exigir dos trabalhadores conhecimento específico sobre qualquer assunto para participar da plataforma, é possível ao solicitante ter acesso também a um subconjunto de trabalhadores especializados, denominados de mestres (*masters*) (Amazon Web Services, 2017). Os mestres são trabalhadores que demonstraram, ao longo do tempo em que participam da plataforma, a capacidade de fornecer bons resultados para tipos específicos de tarefas. Os trabalhadores mestres são selecionados pelo próprio sistema tomando como base o comportamento dos mesmos no sistema, isto é, as estatísticas armazenadas no sistema de reputação. O sistema avalia continuamente os trabalhadores mestres para identificar se os mesmos devem continuar sendo classificados como tal. No entanto, os solicitante podem utilizar os trabalhadores mestres em qualquer HIT desde que essa qualificação seja exigida em seu HIT. Entretanto, não há trabalhadores mestres para todos os tipos de tarefas.

As qualificações no Mturk foram analisadas e classificadas como reputação, padronizadas e customizadas, da seguinte forma:

- **Reputação**: são qualificações mantidas pela plataforma MTurk e que consideram a informação sobre o comportamento do trabalhador no MTurk, isto é, a performance do trabalhador na plataforma. Esta classe de teste de qualificação inclui todos os trabalhadores mestres, incluindo as qualificações denominadas *Masters*, *Categorization masters* e *Photo moderation masters* já incorporadas ao sistema, bem como os tipos de qualificação que levam em consideração estatísticas de desempenho mantidas pela plataforma, tais como HITs submetidos, aprovados, rejeitados e abandonados, usando métricas quantitativas, e gerando informações para serem utilizadas nas seguintes qualificações que são geradas automaticamente pelo Mturk:
  - *HIT submission rate (%)*: Esta qualificação reflete o número de HITs para o qual o trabalhador enviou respostas, dividido pelo número total de HITs aceitos. Sua pontuação é um valor entre 0 e 100. Uma pontuação de 100 indica que o trabalhador enviou uma resposta para cada HIT que aceitou.
  - *HIT approval rate (%)*: Esta qualificação reflete a porcentagem de HITs para os quais o trabalhador submeteu respostas que foram aprovadas, considerando o número total de HITs executados. Sua pontuação é um valor entre 0 e 100. Uma pontuação de 100 indica que o worker enviou uma resposta que foi aprovada para cada HIT que aceitou.
  - *HIT rejection rate (%)*: Esta qualificação reflete o número de HITs para o qual o trabalhador enviou uma resposta que foi rejeitada, dividido pelo número total de HITs, que foram aprovados ou rejeitados. Sua pontuação é um valor entre 0 e 100. Uma pontuação com valor zero indica que nenhum dos HITs submetidos foi rejeitado.
  - *HIT abandonment rate (%)*: qualificação que reflete o número de HITs aceitos pelo trabalhador mas que expiraram antes da resposta ser submetida, dividido pelo número total de HITs aceitos pelo trabalhador. Sua pontuação é um valor entre 0 e 100. Uma pontuação igual a zero indica que o trabalhador não permitiu que nenhum HIT expirasse antes de ter apresentado a resposta.

- *HIT return rate (%)*: Esta qualificação reflete o número de HITs aceitos pelo trabalhador e que em seguida foram devolvidos sem resposta dividido pelo número total de HITs aceitos. Sua pontuação é um valor entre 0 e 100. Uma pontuação de 0 indica que o trabalhador não retornou nenhum HIT.

  Essas informações podem ser utilizadas por qualquer solicitante para selecionar trabalhadores. Por exemplo, se o solicitante não quiser usar trabalhadores recém-registrados no sistema, portanto não têm histórico de trabalho, deve utilizar a qualificação *HIT approval rate* com valor mínimo alto. Logo, manter uma boa reputação no sistema é importante para os trabalhadores porque amplia a possibilidade de realização de mais HITs.

- **Padronizadas**: são todas as qualificações mantidas pela plataforma MTurk e que não pertencem à classe Reputação. Essas qualificações consideram as informações pessoais do trabalhador e não podem ser atualizadas pelo sistema, isto é, são informações do registro do trabalhador que são apenas utilizadas pelo solicitante e verificadas pelo sistema. A atualização é feita exclusivamente pelo próprio trabalhador. Essa qualificação é utilizada quando o solicitante deseja selecionar os trabalhadores utilizando informações de endereço, idade e renda, por exemplo. Por exemplo, se o solicitante deseja selecionar trabalhadores brasileiros ele deve utilizar a qualificação *Location is BR*. Ou se o solicitante deseja utilizar apenas trabalhadores adultos, ele pode exigir a qualificação *Adult Content Qualification*. Nesses casos, nenhuma habilidade específica está sendo exigida do trabalhador.

- **Customizadas**, que permite que o solicitante selecione trabalhadores que possuam as habilidades necessárias para realizar um tipo particular de HIT. Esse tipo de qualificação é criado pelo solicitante.

As qualificações criadas pelos solicitantes são disponibilizadas publicamente permitindo inclusive que um solicitante utilize em seus HIT testes de qualificação criados por outros solicitantes. No entanto, apenas o solicitante que criou o teste de qualificação pode conceder a qualificação ao trabalhador e atribuir a pontuação. O solicitante pode atualizar a pontuação de qualquer trabalhador em qualquer momento. Além disso, cada HIT pode especificar um

limite de pontuação diferente utilizando declaração condicional para indicar o nível de habilidade mínima exigida em uma determinada qualificação para que os trabalhadores possam executar os HITs de cada grupo de HITs específico. Os testes de qualificação são realizados de forma voluntária pelos trabalhadores e não precisam ser remunerados.

As qualificações criadas pelos solicitantes, isto é, do tipo customizadas, podem ser desativadas ou eliminadas, tornando-se inativa. Quando o solicitante desativa uma qualificação, esta não pode ser utilizada nos HITs e, consequentemente, não pode ser solicitada pelos trabalhadores, mas pode ser novamente ativada pelo solicitante, diferentemente de quando é eliminada pelo solicitante. A qualificação também pode se tornar inativa quando deixou de ser utilizada por um período de 120 dias por algum HIT.

O MTurk não fornece ao solicitante um teste que o permita saber se um determinado trabalhador tem ou não determinada qualificação ou todas as qualificações de um determinado trabalhador. O que o solicitante pode fazer é testar se um trabalhador tem determinada qualificação quando esta é adicionada ao seu HIT. Isso é feito durante o projeto da tarefa. A Figura 3.4 apresenta a parte da interface gráfica (seção intitulada *Worker requirements*) em que o solicitante pode fazer suas escolhas. Neste exemplo, o solicitante não marcou a opção de usar trabalhador mestre mas fez uso de 3 tipos de testes de qualificação: o teste do tipo reputação *"Number of HITs Approved"* exigindo valor maior que zero; o teste do tipo padronizado *"Location"* para selecionar trabalhadores que são brasileiros e, teste do tipo customizado *"Qualification test Find 3"* para selecionar trabalhadores que participaram desse teste específico e obtiveram nota igual a 80. Dessa forma, o solicitante escolhe qual teste usar assim com os operadores relacionais e os valores. Através da interface gráfica o solicitante pode fazer uso de até 5 testes de qualificação.

Quanto ao trabalhador, este pode, no entanto, solicitar que lhe seja concedido uma qualificação do tipo customizada. Isso pode ser feito de duas formas. A primeira forma é quando o trabalhador seleciona no *HITs job board* um HIT de seu interesse mas que ele não pode realizar porque não possui a qualificação exigida (ver na Figura 3.3 que cada HIT possui a informação *qualifications required*). Nesse caso, o trabalhador precisa solicitar a permissão de realizar a qualificação exigida. Se conseguir êxito, isto é, obtiver o valor necessário

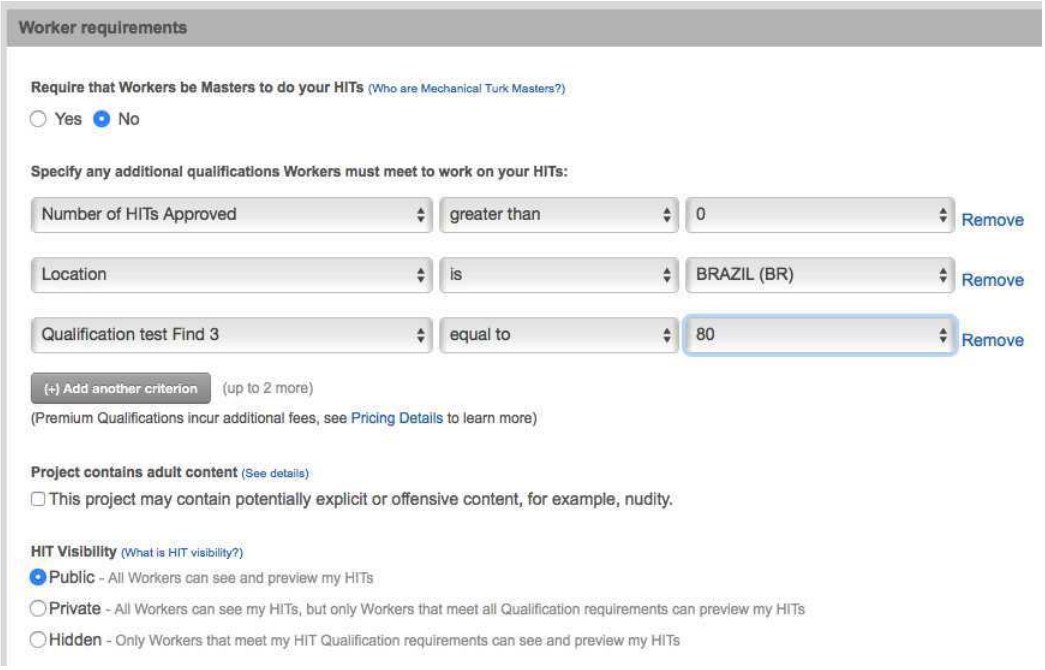


Figura 3.4: Interface para criar tarefas e escolher teste de qualificação no Mturk. Fonte: www.mturk.com.

para a qualificação, os HITs serão liberados para ele. A segunda forma é pesquisando testes de qualificação criados pelos solicitantes disponíveis diretamente na plataforma através do *Qualifications job board* (Figura 3.5), que se torna interessante apenas quando o trabalhador procura se qualificar para HITs de um solicitante ou qualificação conhecidos, pois as qualificações são apresentadas apenas em ordem alfabética, crescente ou decrescente.

Quando a qualificação é do interesse do trabalhador, o mesmo envia uma solicitação ao solicitante que a criou. Ressalta-se que as qualificações que não são criadas pelo solicitante (qualificações padronizadas e de reputação) são apresentadas para os trabalhadores juntamente com as qualificações criadas pelo solicitante, e que aquelas também podem impedir que o trabalhador realize o HIT, caso o trabalhador não atenda ao requisito especificado. O trabalhador, dependendo da qualificação, pode ser direcionado para executar um HIT de teste e o solicitante associará uma pontuação ao seu trabalho. Assim, cada qualificação obtida pelo trabalhador tem uma pontuação associada que é mantida pelo sistema e que determina quais HITs estão disponíveis para o trabalhador realizar. Trabalhadores que alcançam uma pontuação alta o suficiente em uma qualificação podem ser vistos como um grupo de trabalhadores

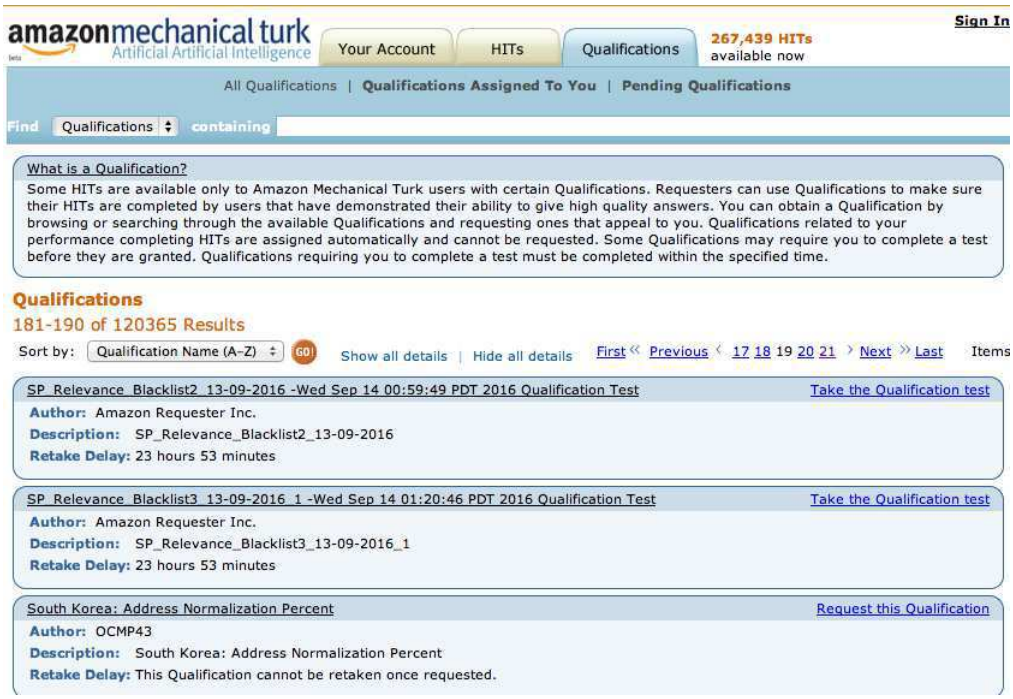


Figura 3.5: Quadro de testes de qualificação do tipo habilidade no MTurk. Fonte: www.mturk.com.

especializados.

3.3 Considerações finais

Este capítulo apresentou estratégias utilizadas com o objetivo de controlar a qualidade dos resultados em plataformas *crowdsourcing*. As estratégias se dividem em dois campos: as que tentam evitar que trabalhadores maliciosos submetam resultados de baixa qualidade e as que analisam os resultados submetidos pelos trabalhadores.

As estratégias apresentadas são mais eficientes para tarefas simples e que possuem um conjunto de respostas possíveis, do que para as tarefas subjetivas que exigem um conhecimento mais específico do trabalhador.

A utilização da reputação do trabalhador é insuficiente no controle da qualidade porque considera apenas o percentual de trabalho realizado que, no contexto de *crowdsourcing* remunerado, pode motivar os trabalhadores a criarem falsas identidades para ludibriar o sis-

tema e assim frustrar o controle de qualidade adotado.

Os mecanismos de prevenção exigem a colaboração do solicitante dado que as habilidades do trabalhador precisam ser testadas em função da tarefa a ser realizada. No entanto, o teste de qualificação parece ser o mais adequado, demandando assim, estudos para investigar sua efetividade.

O funcionamento da plataforma MTurk, assim como os mecanismos de qualificação existentes, também foram explicados nesse capítulo dado que essa foi a plataforma escolhida para os estudos realizados nesse trabalho.

Capítulo 4

Uso de qualificação no MTurk: pesquisa exploratória quantitativa

Existem diversas alternativas para controlar a qualidade dos resultados obtidos em mercados de trabalho on-line ou *crowdsourcing* remunerado, de microtarefas. No entanto, o teste de qualificação, utilizado para fazer uma pré-seleção dos trabalhadores, parece ser o mais adequado, demandando assim, estudos para investigar a sua eficácia.

Testes de qualificação permitem aos solicitantes, além de avaliar a habilidade de um trabalhador na execução de determinada tarefa, treinar o trabalhador ou até mesmo utilizar informações sobre a reputação do trabalhador no sistema ou utilizar as informações fornecidas na ocasião do registro no sistema, como a informação de localização geográfica. Ao utilizar testes de qualificação, portanto, os solicitantes podem direcionar suas tarefas para um grupo específico de trabalhadores.

Entender como os testes de qualificação são utilizados em mercados de trabalho on-line de microtarefas é importante porque, além de suprir a ausência de estudo no tema, fornece subsídios aos solicitantes sobre o comportamento dos trabalhadores em relação às tarefas.

Nesse sentido, este capítulo descreve a pesquisa exploratória quantitativa realizada para ampliar o entendimento de como os solicitantes utilizam testes de qualificação para filtrar trabalhadores em mercado de trabalho on-line de microtarefas.

Este estudo investiga duas questões de pesquisa gerais sobre o quão e como são utilizados os testes de qualificação. A partir desse entendimento, busca-se delinear quais são os tipos de tarefas que usam testes de qualificação.

Os dados analisados nesta pesquisa foram coletados da plataforma Mechanical Turk (MTurk) em dois períodos de tempo distintos.

Nas seções seguintes, detalha-se, inicialmente, os materiais e métodos utilizados, explicando como foi realizada a coleta dos dados e a metodologia utilizada na classificação das tarefas e dos solicitantes. Em seguida, são apresentados os resultados obtidos e as considerações finais.

4.1 Materiais e métodos

4.1.1 Coleta de dados

Foram utilizados dados coletados da plataforma MTurk de dois períodos de tempo distintos. A primeira coleta de dados foi realizada por Ponciano e Brasileiro (2013) por um período de quatro meses (Outubro/2012 a Fevereiro/2013) e, a segunda foi realizada durante um mês no ano de 2016 (Maio/2016), durante a execução desse trabalho.

O processo de coleta de dados foi o mesmo para os dois conjuntos de dados coletados e consistiu em rastrear o sistema e capturar, a cada dois minutos, os dados do quadro de tarefas (Figura 3.3, Seção 3.2.1) e do quadro de qualificações (Figura 3.5, Seção 3.2.2) da plataforma. Do quadro de tarefas, para cada grupo de HITs, foram coletados os seguintes dados: título, solicitante, data de expiração, tempo estimado para a execução, recompensa ofertada, quantidade de HITs disponíveis no grupo de HITs, descrição, palavras-chave e qualificações necessárias. Do quadro de qualificações, para cada qualificação, foram coletados os seguintes dados: título da qualificação, autor, descrição e tempo para refazer a qualificação (*retake delay*).

Uma situação padrão no MTurk é que os HITs são apresentados aos trabalhadores no quadro de tarefas na ordem em que são submetidos ao sistema pelos respectivos solicitantes. Como a taxa de submissão de HITs é alta, em pouco tempo um grupo de HITs desaparece da primeira página do quadro de tarefas. Assim, os grupos de HITs que não são totalmente executados, depois de algum tempo, podem ser removidos e mais tarde submetidos novamente pelo solicitante. Dessa forma, o solicitante consegue colocar o grupo de HITs no topo da lista de HITs disponíveis aumentando a probabilidade de que mais trabalhadores o execute num menor espaço de tempo.

O procedimento de coleta de dados inicialmente adotado não levou em conta a possível redundância de grupos de HITs que são removidos e posteriormente submetidos novamente, uma vez que não existe uma maneira simples de fazer essa filtragem durante a coleta. Assim, os grupos de HITs que são submetidos novamente são considerados como novos HITs na coleta de dados, apesar de já terem sido coletados.

Para o propósito desse estudo, os HITs duplicados foram excluídos. Um processo de filtragem pós-coleta foi aplicado para remover tais duplicatas. Isto foi realizado através da comparação dos HITs coletados, excluindo a janela de tempo, bem como o campo de identificação do HIT durante o processo de comparação. Assim, foram considerados HITs duplicados aqueles que todos os outros campos, ou seja, título, solicitante, data de validade, tempo previsto, recompensa, HITs disponíveis, descrição, palavras-chave e qualificações exigidas, eram iguais a algum HIT previamente coletado.

Em relação às qualificações, todas as qualificações do quadro de qualificações foram coletadas, independente de serem ou não usadas nos HITs coletados em paralelo. Isso significa que nem todas as qualificações coletadas estavam presentes nos HITs coletados. Um processo de filtragem pós-coleta foi adotado com o objetivo de considerar apenas as qualificações utilizadas nos HITs coletados. Além disso, foram consideradas apenas as qualificações que são criadas pelos solicitantes, isto é, do tipo customizada.

4.1.2 Classificação das tarefas

Vários são os tipos de tarefas que são submetidos na plataforma MTurk. Portanto, é importante investigar se o tipo de tarefa tem alguma influência na utilização de testes de qualificação pelos solicitantes.

A classificação das tarefas tem como objetivo identificar grupos de tarefas que são semelhantes entre si e assim permitir uma melhor compreensão da dinâmica de submissão das tarefas em plataformas *crowdsourcing*, assim como uma análise mais detalhada da utilização de testes de qualificação.

As tarefas submetidas no MTurk podem ser classificadas usando seus atributos textuais (por exemplo, título da tarefa, descrição da tarefa e palavras-chave). Assim, nesse trabalho, a classificação das tarefas é realizada de forma similar à classificação de textos, cuja finalidade é a de atribuir categorias predefinidas para documentos com base em seu conteúdo. Dessa forma, cada tarefa é representada pelas palavras contidas nos atributos título e descrição.

Foi utilizada a taxonomia proposta por Gadiraju et al. (2014) para definir essas categorias. Essa taxonomia é um esquema de categorização de dois níveis baseado em um estudo realizado com 1.000 trabalhadores na plataforma CrowdFlower. A partir das respostas coletadas, as classes que descrevem as tarefas típicas da plataforma estudada foram descritas através de um processo manual. As tarefas são classificadas em seis classes orientadas ao objetivo da tarefa, com cada classe contendo subclasses, inclusive de outros tipos de tarefas como apresentado na Tabela 4.1. A categorização de alto nível baseia-se nos objetivos da tarefa, enquanto que as subclasses se baseiam no fluxo de trabalho das tarefas. As classes são explicadas a seguir.

- *Information Finding* (IF) - Tarefas que exigem pesquisa para atender a uma necessidade de informação do solicitante. Por exemplo, uma tarefa que tem como objetivo "encontre o link para fotos de um médico"ou "encontre médicos próximo à cidade de Campina Grande que estão atuando há menos de 2 anos".

- *Verification and Validation* (VV) - Tarefas que exigem que os trabalhadores verifiquem certos aspectos de acordo com algumas instruções passadas pelo solicitante ou confir-

Tabela 4.1: Classes e Subclasses da taxonomia utilizada para tarefas típicas em plataformas *crowdsourcing*

Classe	Subclasse
Information Finding (IF)	Metadata finding
Verification and Validation (VV)	Content Verification, Content Validation, Spam Detection, Data matching
Interpretation and Analysis (IA)	Classification, Categorization, Media Transcription, Data Selection, Sentiment Analysis, Content Moderation, Ranking, Quality Assessment
Content Creation (CC)	Media Transcription, Data Enhancement, Translation, Tagging
Survey (S)	Feedback/Opinions, Demographics
Content Access (CA)	Testing, Promoting

Adaptado de Gadiraju et al. [94]

mem a validade de vários tipos de conteúdo. Exemplos incluem tarefas que solicitam a verificação de transcrições ou verificação, por exemplo, de comportamento spam de algumas contas de usuário.

- *Interpretation and Analysis* (IA) - Trabalhadores utilizam a habilidade de interpretação para executar as tarefas dos solicitantes. Tarefas de classificação se encaixam nessa classe.

- *Content Creation* (CC) - Tarefas que exigem geração de conteúdo. Por exemplo, tarefas do tipo "apresente sugestões de nome para um novo produto"ou "traduza o conteúdo para um determinado idioma".

- *Survey* (S) - Tarefas que envolvem pesquisa sobre uma infinidade de temas, como por exemplo, aspectos comportamentais, satisfação em relação a uma empresa ou produto, aspectos sócio-econômicos, dentre outros.

- *Content Access* (CA) - Trabalhadores devem acessar algum conteúdo, isto é, o trabalhador precisa acessar algum link para realizar a tarefa.

No entanto, dada a grande quantidade de tarefas no conjunto de dados coletados, a classificação automática das tarefas tornou-se necessária. Para isso, utilizou-se o modelo de Máquina de Vetores de Suporte (SVM) (Feldman and Sanger, 2007; Hsu et al., 2016; Yu-Wei, 2015).

Na classificação das tarefas, inicialmente, a abordagem *Bag-Of-Related-Words* (BORW) (Rossi, 2011) foi utilizada para gerar atributos compostos por palavras relacionadas de cada tarefa. A representação BORW é uma extensão da representação *Bag-of-Words* que é bastante utilizada em processamento de linguagem natural e recuperação da informação.

Na representação *Bag-of-Words*, um texto, uma frase ou um documento, é representado como uma sacola (*bag*) de palavras, ignorando a ordem das palavras, informações de pontuação ou informações estruturais. É um modelo comumente usado em métodos de classificação de documentos onde a freqüência de cada palavra é usada como recurso para treinar um classificador.

A representação BORW, por outro lado, tem como objetivo utilizar como atributos palavras relacionadas que se repetem ao longo de uma coleção de documentos. Para extrair as relações entre as palavras são utilizadas regras de associação (Agrawal and Srikant, 1994). Os mecanismos provenientes das regras de associação tornam possível a redução da dimensionalidade, pois, ao invés de utilizar todas as possíveis palavras e combinações de palavras para representar uma tarefa, só são extraídas palavras e relações entre palavras que tenham uma frequência maior que um determinado valor estabelecido.

Uma etapa de pré-processamento foi realizada para tratar e padronizar os dados de entrada de modo que apenas as informações mais relevantes dos atributos título e descrição de cada tarefa fossem preservadas (Feldman and Sanger, 2007). Inicialmente foram removidos os sinais de pontuação, símbolos especiais, números e *stopwords*[1] dos atributos considerados. Em seguida, foram identificadas as palavras sinônimas e realizado o processo de determi-

[1] *Stopwords* são palavras simples de uso cotidiano que não têm relevância no processo de mineração de texto, como por exemplo, os artigos, as preposições, os pronomes, os advérbios e as palavras que são consideradas pouco relevantes para um domínio de aplicação específico.

nação dos radicais dos termos (*stemming*) que faz com que palavras que tenham o mesmo significado, mas que se diferenciam pelo tempo verbal e número, por exemplo, representem a mesma informação. Isso faz com que palavras como *transcript* e *transcription*, por exemplo, correspondam a um único termo e consequentemente sejam representadas e contabilizadas de uma única maneira.

Em seguida, regras de associação foram extraídas das transações. Uma regra de associação caracteriza o quanto a presença de um conjunto de termos numa base de dados implica na presença de algum outro conjunto distinto de itens (Agrawal and Srikant, 1994). Dessa forma, as regras de associação permitem encontrar tendências que podem ser usadas para entender e explorar padrões de comportamento dos dados. Para extrair as regras de associação é necessário, primeiro, gerar o conjunto de itens frequentes (*itemsets*). O algoritmo utilizado para gerar os (*itemsets*) foi o algoritmo *Apriori* (Agrawal and Srikant, 1994). A geração de regras de associação utiliza duas medidas: suporte e confiança. O suporte determina a frequência com que uma regra é aplicável a um determinado conjunto de dados, enquanto a confiança mede a força de implicação descrita pela regra. Os valores considerados para o suporte mínimo e confiança foram, respectivamente, 5% e 75%. Além dos termos simples (unigramas), também foram considerados os bigramas.

Finalmente, no último passo, foi utilizado um modelo de Máquina de Vetores de Suporte (SVM) para classificar automaticamente as tarefas (Feldman and Sanger, 2007; Hsu et al., 2016; Yu-Wei, 2015). Inicialmente, o modelo espaço-vetorial foi utilizado para representar as tarefas e seus atributos. Nesse modelo, cada tarefa é um vetor em um espaço multidimensional, e cada dimensão é um termo identificado (Feldman and Sanger, 2007). Dessa forma, cada tarefa é representada por um vetor $\vec{t} = (v_{i,1}, v_{i,2}, \ldots, v_{i,n})$, onde v_{ij} é um valor binário que indica se o j-ésimo termo está presente ou não na i-ésima tarefa. Em seguida, uma amostra de 600 tarefas foi escolhida aleatoriamente usando a taxonomia apresentada na Tabela 4.1, Seção 4.1.2. Essa amostra foi utilizada para treinar e avaliar o modelo SVM gerado. Os conjuntos de treino e teste foram selecionados usando o método de amostragem estratificada *holdout* (Kohavi, 1995). Este método garante uma distribuição proporcional das diferentes categorias tanto no conjunto de teste como no conjunto de treino.

4.1.3 Classificação dos solicitantes

Vários solicitantes diferentes submetem tarefas na plataforma MTurk. A investigação de como os solicitantes, que têm uma distribuição similar de tarefas quando os diversos tipos são considerados (portfólio), se comportam considerando o uso de teste de qualificação é importante. Para isso, primeiro, os solicitantes foram divididos em grupos com portfólios similares e, em seguida, cada grupo foi analisado individualmente.

A identificação dos grupos de solicitantes nos dados coletados foi realizada através de uma combinação de algoritmos de agrupamento hierárquico e não-hierárquico. Os algoritmos de agrupamento hierárquico possuem a vantagem de serem independentes de parâmetros iniciais e permitem que se investigue uma série de possibilidades de agrupamento. Estes podem ser produzidos através de agrupamento aglomerativo ou por divisão. Os algoritmos de agrupamento não-hierárquicos, por outro lado, otimizam a solução global e fornecem classificações mais robustas a *outliers* do que os hierárquicos. No entanto, a qualidade da classificação depende do número de grupos, que deve ser conhecido *a priori*, assim como dos valores iniciais para os centros dos grupos. A metodologia combina essas duas abordagens usando primeiro o algoritmo de agrupamento hierárquico Ward (Ward, 1963) para explorar uma ampla gama de soluções com diferentes números de agrupamentos. O resultado dessa exploração, em seguida, fornece um número adequado de agrupamentos e seus centros, que são, por sua vez, usados para gerar os centros de grupos do algoritmo não-hierárquico k-means (Hartigan and Wong, 1979). Cada solicitante está presente exatamente em um grupo em nossos resultados.

O portfólio de um solicitante é definido pela porcentagem de tarefas submetidas pelo solicitante em cada uma das seis classes de tarefas consideradas na etapa de classificação das tarefas. Os valores obtidos para cada classe formam os atributos usados pelo algoritmo de agrupamento. Após a execução dos algoritmos de agrupamento, foram identificados sete grupos de solicitantes.

4.2 Apresentação e análise dos resultados

4.2.1 Conjuntos de dados coletados da plataforma MTurk

O resultado da coleta dos dois conjuntos de dados descritos na Seção 4.1.1 é apresentado na Tabela 4.2. Os dados mostram que o número de HITs coletados em um único mês é maior do que o número de HITs coletados em quatro meses. No entanto, quando as duplicatas são eliminadas, os dados sugerem que no conjunto de dados I, mais tarefas foram submetidas mensalmente do que no conjunto de dados II. Isso leva a crer que, provavelmente, os solicitantes, em geral, passaram a utilizar com mais frequência a re-submissão de tarefas para que estas fiquem no topo do quadro de tarefas da plataforma e possam atrair mais trabalhadores. Em relação às qualificações coletadas tem-se que o uso de teste de qualificação customizado foi utilizado pelos solicitantes em ambos os conjuntos de dados, demonstrando que a prática do uso foi continuada com o passar do tempo.

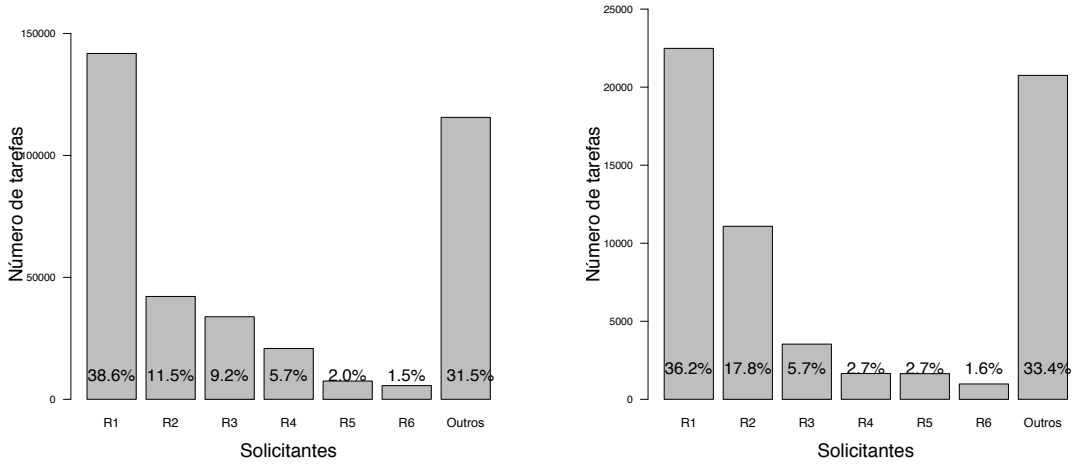
Tabela 4.2: Resultado das coletas de dados.

	Conjunto de dados I (Período da coleta: 4 meses)	Conjunto de dados II (Período da coleta: 1 mês)
Número de HITs coletados	2,5 milhões	3,4 milhões
Número de HITs considerados	367.413	62.141
Total de qualificações válidas	1.192	1.951
Qualificação do tipo reputação	8	7
Qualificação do tipo padronizada	5	2
Qualificação do tipo customizada	1.179	1.942

4.2.2 Os solicitantes

O conjunto de solicitantes foi analisado com o objetivo de identificar a quantidade de solicitantes diferentes que submeteram tarefas no período em que os dados foram coletados. Para o conjunto de dados I, foram identificados 8.019 solicitantes distintos enquanto que, para o conjunto de dados II, foram identificados 4.142 solicitantes distintos.

Uma inspeção preliminar, considerando a quantidade de tarefas submetidas pelos diferentes solicitantes, nos dois conjuntos de dados, revela uma forte concentração de tarefas submetidas por poucos solicitantes ativos e muitos solicitantes que submetem poucas tarefas, uma tendência também identificada por Ipeirotis (2010b).



(a) Solicitantes do conjunto I

(b) Solicitantes do conjunto II

Figura 4.1: Perfil dos solicitantes considerando o número de tarefas submetidas

A Figura 4.1 apresenta o número de tarefas submetidas pelos seis solicitantes mais ativos e por todos os outros solicitantes. Para o conjunto de dados I, os seis solicitantes foram responsáveis por 68,5% de todas as submissões e, apenas os solicitantes R1 e R2, juntos, por 50,1% (Figura 4.1a). Por outro lado, a Figura 4.1b mostra que, no conjunto de dados II, os seis solicitantes foram responsáveis por 66,6% das submissões, enquanto que R1 e R2, juntos, submeteram 54% de todas as tarefas. Logo, os dois conjuntos de dados apresentam comportamento similar em relação à submissão de tarefas, isto é, uma distribuição enviesada.

A distribuição enviesada dos dados pode provocar um viés tanto na distribuição do uso de testes de qualificação como também na distribuição das tarefas nas classes. Por esse motivo, para as análises, os dados são apresentados de duas maneiras: i) considerando todas as tarefas coletadas; e, ii) considerando apenas as tarefas apresentadas pelos demais solicitantes, isto é, considerando apenas as tarefas submetidas por todos os solicitantes, com exceção dos

solicitantes R1 e R2 (nesse caso, os dados são apresentados entre parênteses e referenciados como sendo dados relativos à cauda da distribuição).

4.2.3 Classificação das tarefas

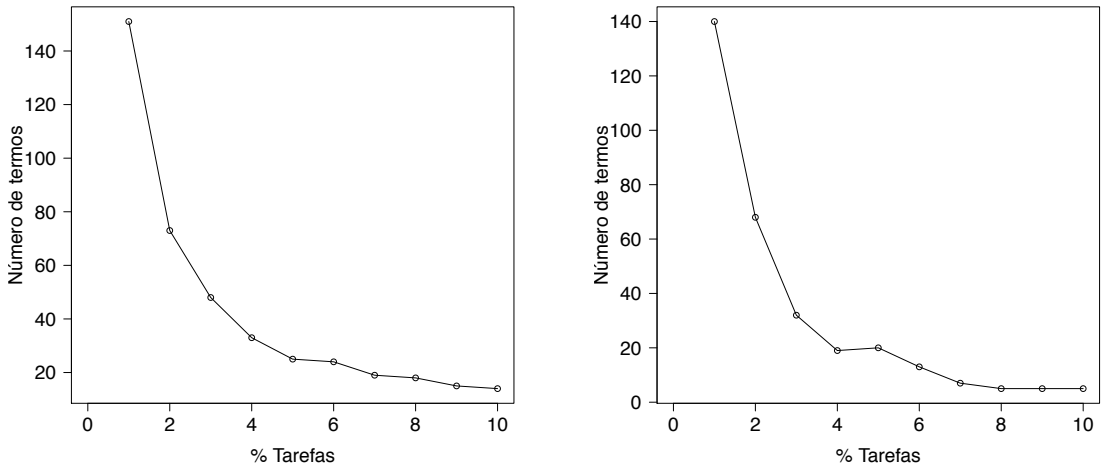
O processamento de classificação iniciou com a etapa de pré-processamento com o intuito de tratar e padronizar os termos contidos nos campos título e descrição de cada tarefa. A Tabela 4.3 apresenta alguns exemplos de tarefas do Conjunto de Dados I (título e descrição) e a sua representação através dos seus respectivos termos após a padronização.

Tabela 4.3: Exemplo da etapa de pré-processamento dos campos das tarefas.

Título	Descrição	Representação após pré-processamento
Transcription Review A534636 (audio length: 4 minutes 2 seconds)	Review an audio file transcription to find errors and correct them	a534636 audio correct error file find length minut review second transcript
1000 Word Article: LG Appliances (Brand Overview and Appliance Repair)	Write an informative 1000 word article	applianc articl brand inform lg overview repair word write
Edit: 350 Words of Content	Review and make basic edits to 350 words for an online retailer	basic content edit onlin retail review word
Find Wikipedia articles corresponding to short text fragments	This task involves finding Wikipedia articles which correspond to	articl correspond find fragment involv short task text wikipedia
Image Search using Descriptions (Plus BONUS!)	You will perform image search by describing an image in different ways. You will get a small BONUS (up to 20 cents) depending on the quality of your descriptions. The whole HIT should take about than 20 min.	bonu cent depend describ descript hit imag min perform qualiti search small way

Após o processo de pré-processamento verificou-se que o número de termos distintos identificados foi de 43.210 e 17.299, para o conjunto de dados I e II, respectivamente. Pelos exemplos apresentados na Tabela 4.3 percebe-se que alguns termos não contribuem para a classificação, como por exemplo, *a534636*, *applianc* e *lg*. Para reduzir a quantidade de termos e viabilizar o processo de extração de regras, a distribuição dos termos em relação às tarefas foi analisada. Os dados revelaram que uma porcentagem significativa dos termos identificados (99,65% e 99%, para o conjunto de dados I e II, respectivamente) aparece em uma porcentagem muito baixa de tarefas (menos de 1%), portanto, esses termos não são apropriados para classificar as tarefas. A partir da análise do número de termos que aparecem em pelo menos um certo percentual de tarefas (Figura 4.2), arbitrou-se que apenas os termos que aparecessem em pelo menos 5% e 6% das tarefas, respectivamente no conjunto de dados I e II, seriam escolhidos para a classificação das tarefas. Dessa forma, 25 e 19 termos foram considerados, respectivamente, para gerar as transações associadas a cada tarefa de cada

conjunto de dados contendo apenas os termos significativos a ela associados e dispostos em ordem lexicográfica.



(a) Conjunto de dados I

(b) Conjunto de dados II

Figura 4.2: Número de termos que aparecem em pelo menos x% das tarefas

A Tabela 4.4 apresenta os 25 termos considerados pelo algoritmo de classificação e os bigramas, juntamente com o suporte, para o conjunto de dados I. Uma análise preliminar da matriz binária atributo-valor gerada para o processo de classificação permitiu concluir que alguns termos não acrescentavam valor à classificação por serem muito genéricos. Dessa forma, os termos destacados na tabela foram removidos da matriz. O mesmo procedimento foi realizado para o conjunto de dados II.

Geradas as transações, uma amostra de 600 tarefas foi escolhida aleatoriamente com o objetivo de treinar e avaliar o modelo SVM. As tarefas da amostra foram classificadas manualmente. O avaliador analisou os atributos título e descrição e usou a taxonomia para selecionar a classe mais adequada para cada tarefa.

Para verificar a qualidade da classificação, foi selecionada uma amostra menor de tarefas para ser classificada manualmente por uma outra pessoa. Inicialmente foram selecionadas 100 tarefas aleatoriamente para serem classificadas por um humano. As tarefas foram analisadas e aquelas cujo título e descrição estavam repetidos foram eliminadas da amostra. Dessa

Tabela 4.4: Termos relevantes (unigramas e bigramas) usados na classificação e o suporte associado do Conjunto de Dados I.

Termo	**Suporte (%)**	**Termo**	**Suporte (%)**	**Termo**	**Suporte (%)**
transcript	54,35	**length**	11,50	question	5,21
audio	49,54	text	11,29	audio, transcript	49,46
bonu	41,28	word	9,73	**style, guid**	35,87
style	36,13	articl	8,84	**style, audio**	35,86
guid	35,91	find	8,21	guid, transcript	35,86
mmss	35,86	difficult	8,20	. style, transcript	35,86
avg.	35,77	provid	7,79	audio, guid	35,86
easi	26,51	write	6,96	text, transcript	10,17
premium	24,94	review	6,64	audio, text	8,47
express	18,18	imag	6,45	word, articl	6,20
minut	14,58	**hit**	6,08	articl, write	5,33
second	12,65	record	6,00		

forma, a amostra resultante ficou com 50 tarefas distribuídas da seguinte forma: 7 da classe IF, 4 da classe VV, 9 da classe IA, 8 da classe S, 7 da classe None e 15 da classe CC. O grau de concordância, isto é, quando a classificação foi idêntica à primeira classificação da amostra, foi de 88%. Apenas seis tarefas foram classificadas de forma diferente. No entanto, as diferenças nas classificações ocorreram em função do entendimento do texto, principalmente quando o texto da descrição possuía mais de uma ação deixando o classificador em dúvida sobre qual seria a classe mais adequada. Por exemplo, a descrição *"Please choose the best answer(s) for the Katters Australian Party of Australia position on Nuclear Waste"* pode ser entendida como ação de gerar conteúdo (classe CC) ou como ação de interpretar e analisar (classe IA). Já na descrição *"Given information, find the closest match/corresponding row in a spreadsheet"*, por exemplo, o entendimento pode ser de buscar a informação (classe IF) ou gerar conteúdo (classe CC). O fato de que em certos casos é possível que a tarefa pertença a mais de uma classe também foi constatado por Gadiraju (2014), autor da taxonomia utilizada na classificação.

Da amostra de 600 tarefas, 70% das tarefas foi utilizada para treinamento e 30% para testes, para cada classe, para cada conjunto de dados, conforme é apresentado na Tabela 4.5. A amostra da classe de tarefas CC é a que tem o maior número de tarefas em ambos os conjuntos de dados indicando que solicitantes submetem mais tarefas dessa classe.

Tabela 4.5: Número de tarefas no conjunto de treino e de teste de cada conjunto de dados.

	Conjunto de dados I			**Conjunto de dados II**		
Classe	**#Tarefas**	**Treino**	**Teste**	**#Tarefas**	**Treino**	**Teste**
Information Finding (IF)	45	31	14	17	12	5
Verification and Validation (VV)	49	34	15	108	76	32
Interpretation and Analysis (IA)	33	24	9	51	36	15
Content Creation (CC)	355	249	106	261	183	78
Survey (S)	29	20	9	66	46	20
None (N)	89	62	27	97	68	29
Total	600	420	180	600	420	180

Ressalta-se que nenhuma tarefa foi classificada como sendo da classe *Content Access*. A primeira suspeita foi de que tarefas dessa classe haviam sido pouco submetidas, e por isso, com baixa probabilidade de serem selecionadas na amostra gerada. No entanto, essa suspeita

foi eliminada quando os conjuntos de dados foram analisados manualmente, fazendo buscas por termos relativos a esse tipo de tarefa e nenhuma tarefa nessa classe ter sido encontrada. Logo, pode-se concluir que este tipo de tarefa não estava presente nos conjuntos de dados coletados. Além disso, algumas tarefas não se enquadravam em nenhuma das 6 classes descritas na Tabela 4.1, Seção 4.1.2. Essas tarefas foram classificadas nesse trabalho como sendo de uma nova classe denominada *None*.

Para quantificar o desempenho da classificação automática realizada, a matriz de confusão foi gerada para ambos os conjuntos de dados (Tabelas 4.6 e 4.7). Nessas matrizes, as colunas representam as tarefas que foram classificadas manualmente, enquanto as linhas representam as tarefas que foram classificadas pelo modelo SVM. Os maiores valores na matriz confusão situam-se na diagonal principal, o que significa que a maioria das tarefas foi corretamente classificada pelo modelo SVM para os dois conjuntos de dados. No geral, a precisão média de classificação do modelo SVM foi de 98,3% e 85.5% e o coeficiente Kappa foi de 0,97 e 0,79, respectivamente, para os conjuntos de dados I e II. Uma vez que os valores acima de 0,75 são considerados como tendo um alto nível de concordância (Carletta, 1996), podemos considerar que o modelo SVM produziu excelentes resultados.

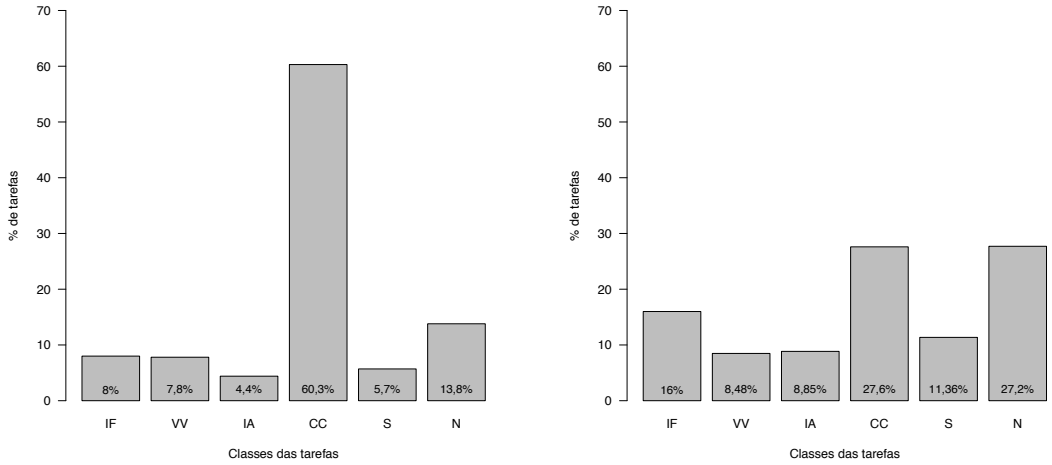
Tabela 4.6: Matriz confusão do modelo SVM para o conjunto de dados I.

	IF	**VV**	**IA**	**CC**	**S**	**N**	Acurácia da classificação do modelo SVM
IF	**14**	0	1	0	0	0	93.3%
VV	0	**14**	0	0	0	0	100.0%
IA	0	0	**8**	0	0	0	100.0%
CC	0	1	0	**106**	1	0	98.1%
S	0	0	0	0	**8**	0	100.0%
N	0	0	0	0	0	**27**	100.0%

Tabela 4.7: Matriz confusão do modelo SVM para o conjunto de dados II.

	IF	**VV**	**IA**	**CC**	**S**	**N**	Acurácia da classificação do modelo SVM
IF	**2**	0	0	0	0	0	100.0%
VV	0	**32**	0	0	0	0	100.0%
IA	1	0	**6**	0	0	0	85.7%
CC	1	0	6	**78**	4	10	78.8%
S	0	0	0	0	**16**	0	100.0%
N	1	0	3	0	0	**19**	82.6%

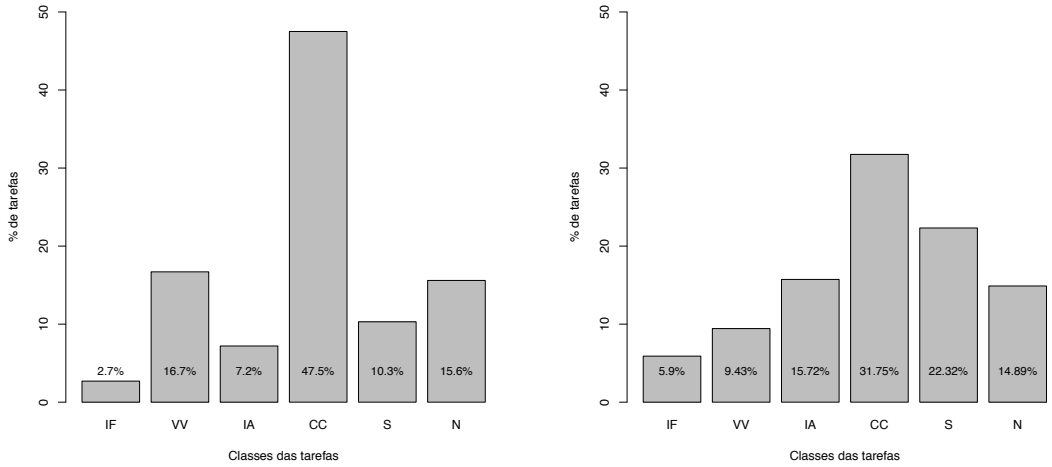
A distribuição das tarefas nas diferentes classes é apresentada na Figura 4.3 para o conjunto de dados I e na Figura 4.4 para o conjunto de dados II.



(a) Todas as tarefas

(b) Apenas as tarefas submetidas pelos solicitantes da cauda da distribuição

Figura 4.3: Distribuição das tarefas nas classes consideradas para o conjunto de dados I.



(a) Todas as tarefas

(b) Apenas as tarefas submetidas pelos solicitantes da cauda da distribuição

Figura 4.4: Distribuição das tarefas nas classes consideradas para o conjunto de dados II.

Para o conjunto de dados I, quando todas as tarefas são consideradas (Figura 4.3a), a classe *Content Creation* (CC) apresenta a maior porcentagem de tarefas submetidas (60,3%). A porcentagem de tarefas que não se enquadrava em nenhuma das classes definidas pela taxonomia utilizada foi de 13,8%. Ao considerar apenas as tarefas apresentadas pelos solicitantes da cauda da distribuição (Figura 4.3b), a classe CC ainda responde por uma grande porcentagem das tarefas (27,6%), mas não tão distintamente como quando todo o conjunto de dados é considerado. Isso ocorre porque quase todas as tarefas submetidas pelo solicitante R1 e, mais de 2/3 das tarefas submetidas pelo solicitante R2 são da classe CC. A remoção dessas tarefas, basicamente, duplicou a proporção de tarefas nas classes *Information Finding* (IF), *Interpretation and Analysis* (IA) e *Survey* (S), bem como na classe N. As tarefas da classe *Verification and Validation* (VV) não tiveram uma alteração importante porque apenas 30% tarefas enviadas por R2 estão nessa classe.

Considerando o conjunto de dados II, quando todas as tarefas são consideradas (Figura 4.4a), a classe *Content Creation* (CC) apresenta a maior porcentagem de tarefas submetidas (47,5%). A porcentagem de tarefas que não se enquadrava em nenhuma das classes definidas pela taxonomia utilizada foi de 15,6%. Ao considerar apenas as tarefas apresentadas pelos solicitantes da cauda da distribuição (Figura 4.4b), a classe CC ainda responde por uma grande porcentagem das tarefas (31,75%), mas não tão distintamente como quando todo o conjunto de dados é considerado. Isso ocorre porque mais de 63% das tarefas submetidas por R1 e 57% das tarefas submetidas por R2 são da classe CC. A remoção dessas tarefas, basicamente, aumentou a proporção de tarefas nas classes *Information Finding* (IF), *Interpretation and Analysis* (IA) e *Survey* (S) em mais do que o dobro. As tarefas da classe *Verification and Validation* (VV) e *None* (N) tiveram pequena redução decorrente do fato de que as demais tarefas submetidas por R1 e R2, respectivamente, serem daquelas classes.

Em suma, os dados demonstram que nos dois conjuntos de dados a classe de tarefas com maior número de tarefas é a classe CC e que os solicitantes R1 e R2 influenciam os resultados das classes CC e VV.

4.2.4 Distribuição do uso de testes de qualificação

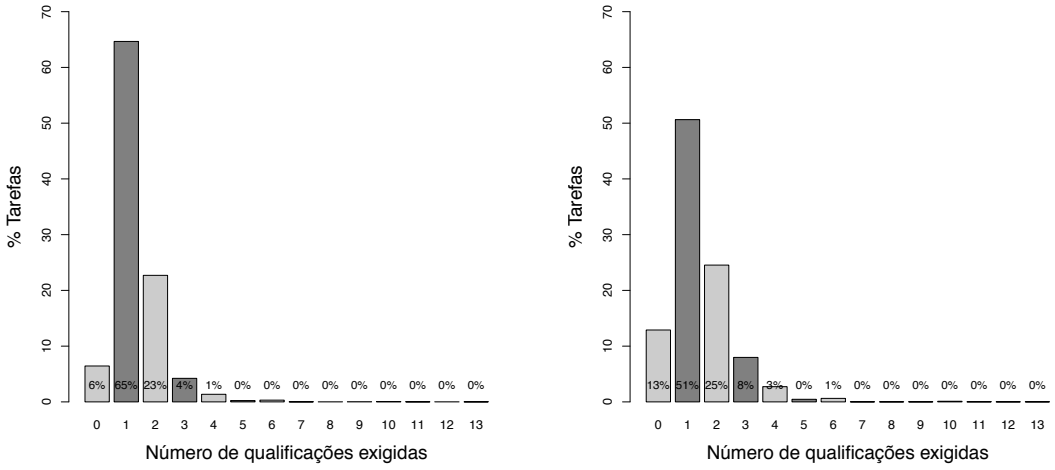
Esta seção descreve como os solicitantes usam os testes de qualificação no MTurk, considerando os três tipos de qualificações descritos na Seção 3.3.2: reputação, padronizada e customizada. O objetivo é responder as seguintes questões:

- Qual a proporção de tarefas que usam testes de qualificação?
- Quão diversa é a quantidade de testes de qualificação usada nas tarefas?
- Qual é o teste de qualificação mais comumente utilizado pelos solicitantes?

Entende-se que ao responder tais questões seja possível caracterizar o uso dos requisitos de qualificação a partir da perspectiva das diferentes classes de tarefas.

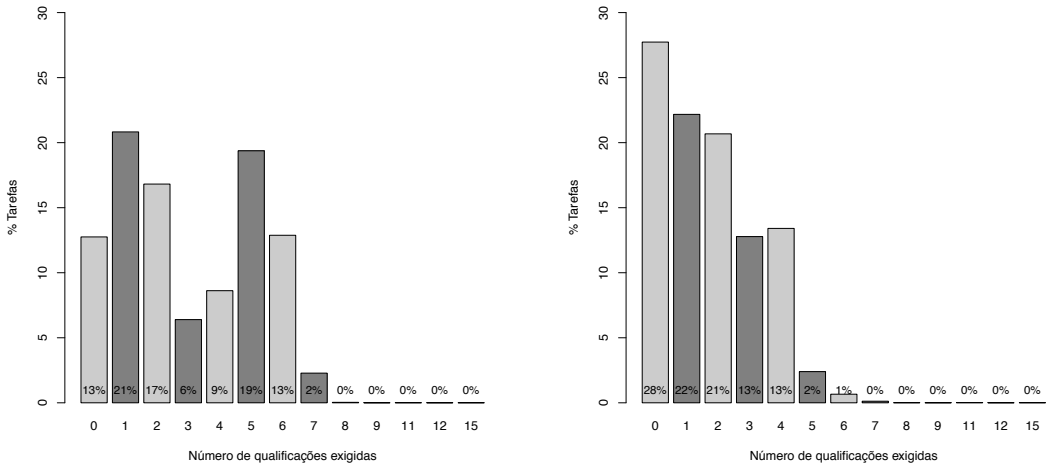
A variação no número de testes de qualificação presente em uma única tarefa é diferente para os conjuntos de dados analisados. Para o conjunto de dados I, o intervalo do número de testes de qualificação presente em uma única tarefa varia de 0 a 13. Das 367.413 (183.421) tarefas coletadas apenas 6% (13%) não exigiram nenhum tipo de teste de qualificação, 64,7% (50,6%) exigiu apenas uma única qualificação, 22,7% (24,5%) exigiu duas, 4,2% (8%) exigiu três, 2% (4%) exigiu quatro ou mais qualificações. A Figura 4.5 mostra a porcentagem de tarefas que utiliza um número específico de qualificações. Considerando apenas as tarefas que utilizam teste de qualificação, 97,9% (95,5%) usa no máximo três tipos de testes. Além disso, cerca de 2/3 das tarefas usa apenas um único tipo de qualificação em ambas as distribuições.

Por outro lado, para o conjunto de dados II, o intervalo do número de testes de qualificação presente em uma única tarefa varia de 0 a 15. Das 62.141 (28.752) tarefas coletadas apenas 13% (28%) não exigiram nenhum tipo de teste de qualificação, 21% (22%) exigiu apenas uma qualificação, 17% (21%) exigiu duas, 6% (13%) exigiu três, 9% (13%) exigiu quatro, 19% (2%) exigiu cinco, 13% (1%) exigiu seis, e 2% (0,2%) exigiu sete ou mais qualificações. A Figura 4.6 mostra a porcentagem de tarefas que utiliza um número específico de qualificações.



(a) Todas as tarefas (b) Apenas a cauda da distribuição

Figura 4.5: Porcentagem de tarefas que usam um número particular de qualificações no conjunto de dados I.



(a) Todas as tarefas (b) Apenas a cauda da distribuição

Figura 4.6: Porcentagem de tarefas que usam um número particular de qualificações no conjunto de dados II.

A Figura 4.6a mostra que há um percentual de tarefas (32%) que utiliza cinco ou seis qualificações. Logo, 85% das tarefas submetem com até 6 qualificações. Considerando apenas as tarefas que usam teste de qualificação, 97,3% não requerem mais do que seis qualificações. Quando os solicitantes mais ativos, R1 e R2, não são considerados na análise

(Figura 4.6b), 69% das tarefas utiliza até quatro testes de qualificação (95,5% não requerem mais do que quatro qualificações, considerando apenas as tarefas que usam qualificação).

Essa alta concentração, em ambos os conjuntos de dados, justifica, portanto, analisar como os solicitantes usam os diferentes tipos de qualificações em suas tarefas considerando apenas o uso de até três ou até seis testes de qualificação. As Tabelas 4.8 e 4.9 apresentam um resumo da distribuição de como as combinações dos testes de qualificação são usados nas tarefas para os conjuntos de dados I (até três qualificações) e II (até seis qualificações) para pré-selecionar os trabalhadores. As combinações cujos percentuais são iguais a zero foram omitidas da tabela.

Tabela 4.8: Resumo da distribuição dos testes de qualificação utilizados nas tarefas do conjunto de dados I.

Número de qualificações exigidas na tarefa	**Teste de qualificação**			
	Customizado	**Reputação**	**Padronizado**	**% consid. tudo (% consid. cauda)**
1	1	0	0	78,3 (44,6)
	0	1	0	10,9 (27,8)
	0	0	1	10,8 (27,6)
2	2	0	0	51,0 (9,1)
	0	2	0	12,7 (23,5)
	0	0	2	2,2 (4,2)
	1	1	0	8,1 (15,1)
	1	0	1	2,9 (5,4)
	0	1	1	23,1 (42,8)
3	3	0	0	4,3 (1,0)
	0	3	0	6,1 (6,5)
	0	0	3	1,5 (1,6)
	2	1	0	0,6 (0,7)
	2	0	1	0,3 (0,3)
	0	2	1	74,6 (76,7)
	0	1	2	1,7 (1,8)
	1	2	0	5,6 (6,0)
	1	0	2	0,0 (0,0)
	1	1	1	5,1 (5,4)

Quando considera-se todas as tarefas do conjunto de dados, verifica-se a predominância da presença do tipo de qualificação customizada, no conjunto de dados I, quando somente uma

Tabela 4.9: Resumo da distribuição dos testes de qualificação usados nas tarefas do conjunto de dados II.

Número de qualificações exigidas na tarefa	**Teste de qualificação**			
	Customizado	**Reputação**	**Padronizado**	**% consid. tudo (% consid. cauda)**
1	1	0	0	83 (65)
	0	1	0	9 (18)
	0	0	1	8 (16)
2	2	0	0	49 (9)
	0	2	0	11 (19)
	0	0	2	0 (1)
	1	1	0	1 (3)
	1	0	1	16 (11)
	0	1	1	23 (40)
3	3	0	0	9 (0)
	0	3	0	5 (5)
	2	1	0	4 (4)
	2	0	1	9 (10)
	0	2	1	53 (58)
	0	1	2	1(1)
	1	2	0	4 (4)
	1	0	2	2 (2)
	1	1	1	14 (15)
4	4	0	0	29 (0)
	3	0	1	10 (14)
	0	3	1	2 (3)
	2	2	0	1 (1)
	2	0	2	1(2)
	0	2	2	28 (39)
	1	2	1	28 (40)
	2	1	1	1 (1)
5	5	0	0	94 (0)
6	6	0	0	98 (0)

única (78,3%) ou duas (51,0%) qualificações são utilizadas, enquanto que no conjunto de dados II, quando somente uma (83%), duas (49%), cinco (94%) ou seis (98%) qualificações são utilizadas. Além disso, no conjunto de dados II, os dados absolutos para todas as demais possibilidades de combinações com cinco ou seis qualificações são exatamente iguais para todo o conjunto de dados e a cauda. O solicitante R1 é o único responsável pelo uso de qualificações customizadas quando o número de qualificações é cinco ou seis. Por isso, quando analisa-se a cauda o percentual equivalente é igual a zero.

Quando três qualificações são exigidas, então a presença das qualificações reputação e padronizada é bem maior se comparada com a presença da qualificação customizada, tanto para o conjunto de dados I (93,8%, 87,4% e 16%, para reputação, padronizada e customizada, respectivamente) como para o conjunto de dados II (80%, 79% e 41%, para reputação, padronizada e customizada, respectivamente). Porém, a combinação do uso de duas qualificações (reputação associada com padronizada) é a mais frequente, considerando todos os dados ou apenas a cauda, em ambos os conjuntos de dados.

Para o conjunto de dados II, quando quatro qualificações são solicitadas nas tarefas, o comportamento da cauda é diferente do conjunto total porque não são consideradas as tarefas submetidas pelo solicitante R1. Logo, as qualificações do tipo reputação e padronizada prevalecem em relação à customizada. Considerando-se todo o conjunto de dados, as qualificações parecem ser usadas de forma parecida, isto é, 70%, 60% e 71% das tarefas usam, respectivamente, qualificações do tipo customizada, reputação e padronizada.

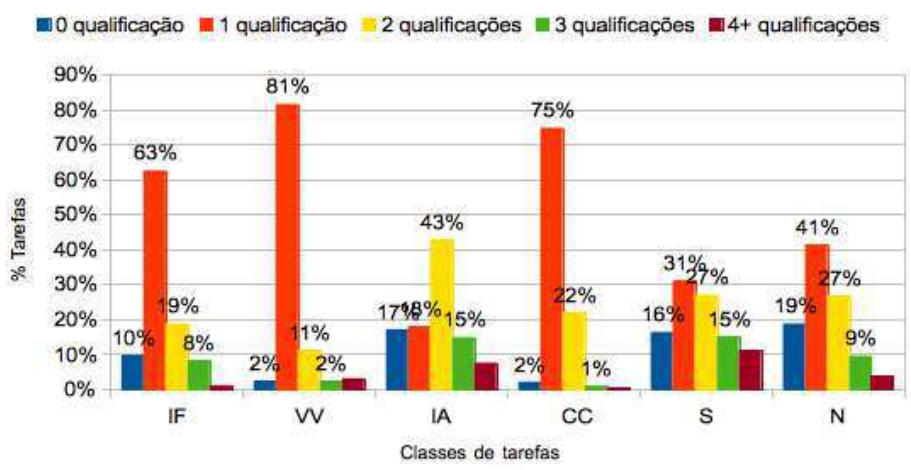
Observa-se, portanto, que quando o número de qualificações utilizado nas tarefas aumenta (considerando até quatro qualificações), as qualificações do tipo reputação e padronizada são mais utilizadas, o que sugere a necessidade de um número maior de qualificações mais simples para construir um mecanismo de pré-seleção que capture com mais precisão as necessidades dos solicitantes.

A Figura 4.7 apresenta a distribuição do uso de teste de qualificação nas diferentes classes para o conjunto de dados I. A Figura 4.7a apresenta os dados considerando todas as tarefas coletadas. Em todas as classes, há uma predominância de tarefas que usam pelo menos um

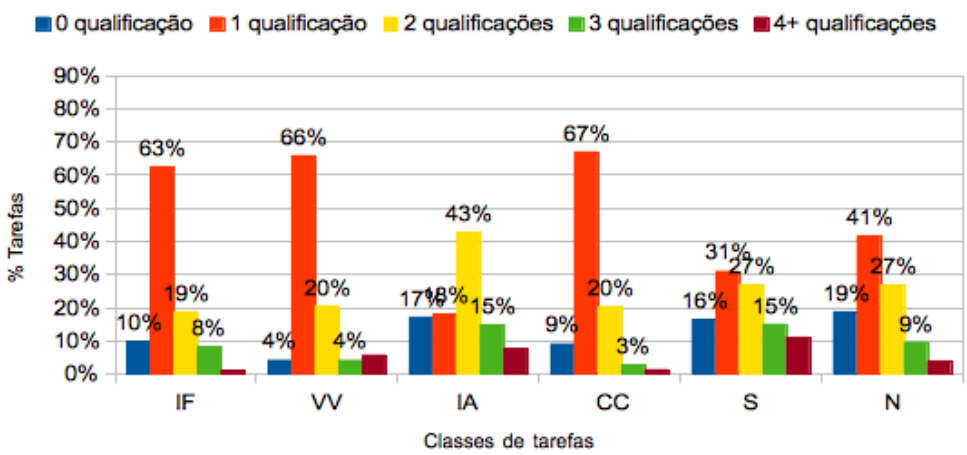
teste de qualificação (não inferior a 81%), sendo a classe *Content Creation* (CC) a classe com a menor porcentagem de tarefas sem mecanismo de pré-seleção de trabalhador (2,1%), e a classe *None* (N) com a maior porcentagem dessas tarefas (18,7%). A porcentagem de tarefas usando uma, duas ou três qualificações é variável entre as diferentes classes. Para as classes *Information Finding* (IF), *Verification and Validation* (VV) e CC existe uma clara predominância do uso de uma única qualificação, enquanto para a classe *Interpretation and Analysis* (IA) o mais comum é o uso de duas qualificações. As classes *Survey* (S) e *None* (N) apresentam maior quantidade de tarefas usando um e dois testes de qualificações. Também, para S e IA, uma porcentagem relativamente menor de tarefas usam quatro ou mais qualificações. Como esperado, quando as tarefas submetidas pelos solicitantes R1 e R2 (Figura 4.7b são desconsideradas, as únicas classes que são fortemente impactadas são CC e VV, devido ao fato de que estes solicitantes submetem principalmente tarefas dessas classes, sendo mais frequente a submissão de tarefas da classe CC.

Para o conjunto de dados II, a distribuição do uso de teste de qualificação nas diferentes classes analisadas é apresentada na Figura 4.8. Quando todos os dados são considerados (Figura 4.8a) em todas as classes, há uma predominância de tarefas que usam pelo menos um requisito de qualificação (não inferior a 65%), sendo a classe *Verification and Validation* (VV) a classe com a menor porcentagem de tarefas sem mecanismo de pré-seleção de trabalhador (2%), e a classe *Information Finding* (IF) com a maior porcentagem dessas tarefas (35%). A porcentagem de tarefas usando de uma a sete ou mais qualificações é variável entre as diferentes classes. As classes CC e IA apresentam uma predominância do uso de uma única qualificação assim como de cinco e quatro qualificações, respectivamente. A classe VV apresenta uma porcentagem maior de tarefas com cinco qualificações enquanto que para as classes S e N o mais comum é o uso de duas qualificações. A classe IF contém um pouco mais de tarefas com três qualificações do que com uma ou quatro qualificações. Como esperado, quando as tarefas submetidas pelos solicitantes R1 e R2 são ignoradas (Figura 4.8b), as classes que são fortemente impactadas são CC, VV e N, devido ao fato de que estes solicitantes submetem principalmente tarefas dessas classes, sendo mais frequente a submissão de tarefas da classe CC.

A distribuição do uso das qualificações por tipo (reputação, padronizada e customizada)



(a) Todas as tarefas

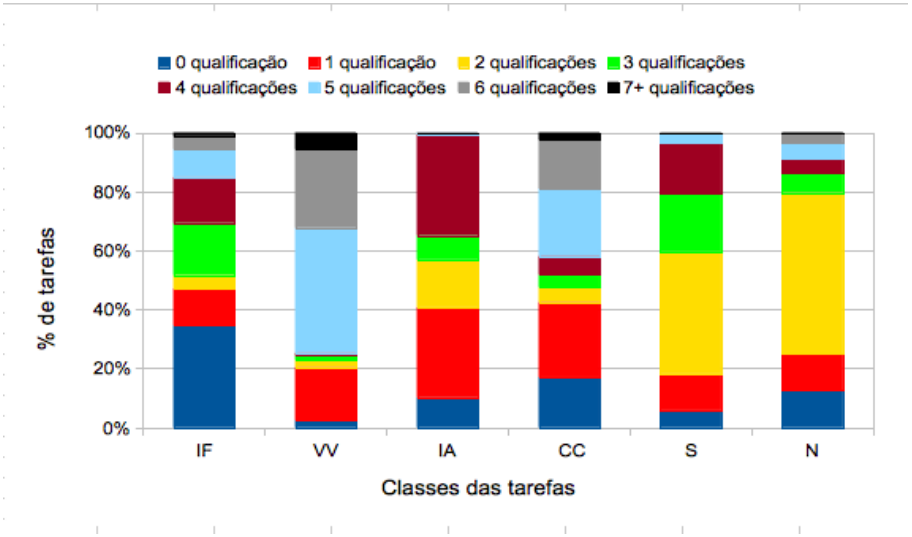


(b) Apenas a cauda da distribuição

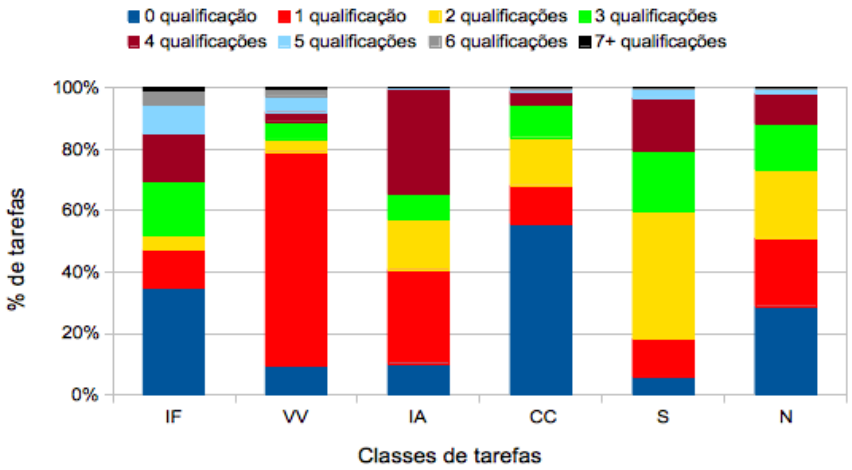
Figura 4.7: Distribuição do número de qualificações solicitadas nas diferentes classes de tarefas do conjunto de dados I.

para cada classe de tarefas, considerando os conjuntos de dados I e II, é apresentada nas Figuras 4.9 e 4.10, respectivamente.

No conjunto de dados I, quando todo o conjunto de dados é analisado (Figura 4.9a), os resultados mostram que há uma variação considerável na distribuição dos testes de qualificações nas diferentes classes. As classe IA e S têm distribuições similares, apresentando predominância do teste de qualificação do tipo reputação e padronizada, especialmente con-



(a) Todas as tarefas

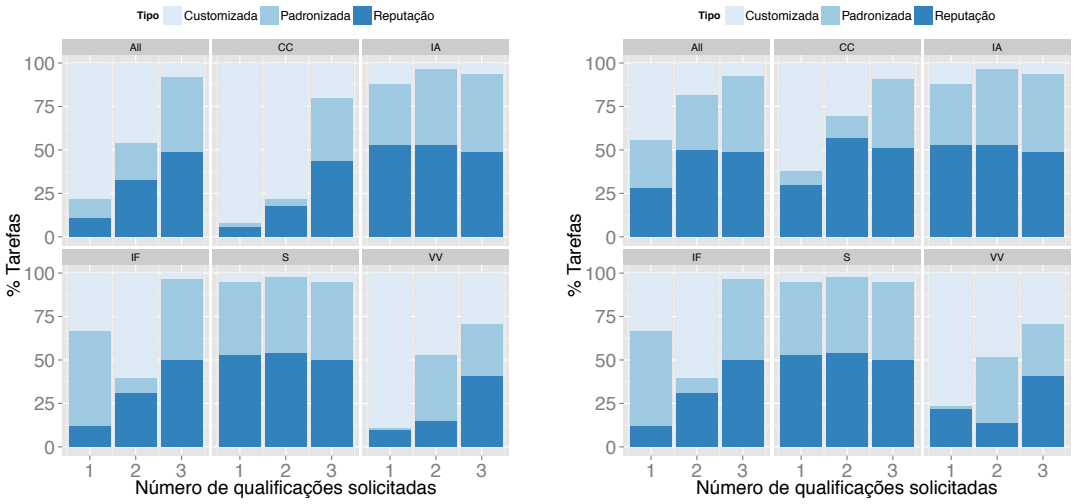


(b) Apenas a cauda da distribuição

Figura 4.8: Distribuição do número de qualificações solicitadas nas diferentes classes de tarefas do conjunto de dados II.

siderando que para essas classes, tarefas com uma única qualificação não são prevalentes. As classes CC e VV também têm distribuições similares, especialmente para os casos de tarefas que usam uma e três qualificações. No caso de uma única exigência de qualificação, as qualificações customizadas predominam. Por outro lado, as tarefas com dois testes de qualificação são duas vezes mais frequentes em CC do que em VV. Também vale a pena ressaltar que os resultados globais (*All*) também são semelhantes aos resultados apresentados para as classes CC e VV. Isso ocorre porque os resultados gerais são fortemente afetados pelo fato de

que as tarefas de CC são muito mais populares do que qualquer outra classe de tarefa (vide Figura 4.3). Finalmente, a classe IF apresenta uma distribuição diferente de todas as outras classes, sendo a qualificação do tipo padronizada predominante para o caso quando um único teste de qualificação é usado, a qualificação do tipo customizada predominante quando são usados dois testes de qualificação e, reputação e padronizada predominante quando três testes de qualificação são usados.



(a) Todas as tarefas

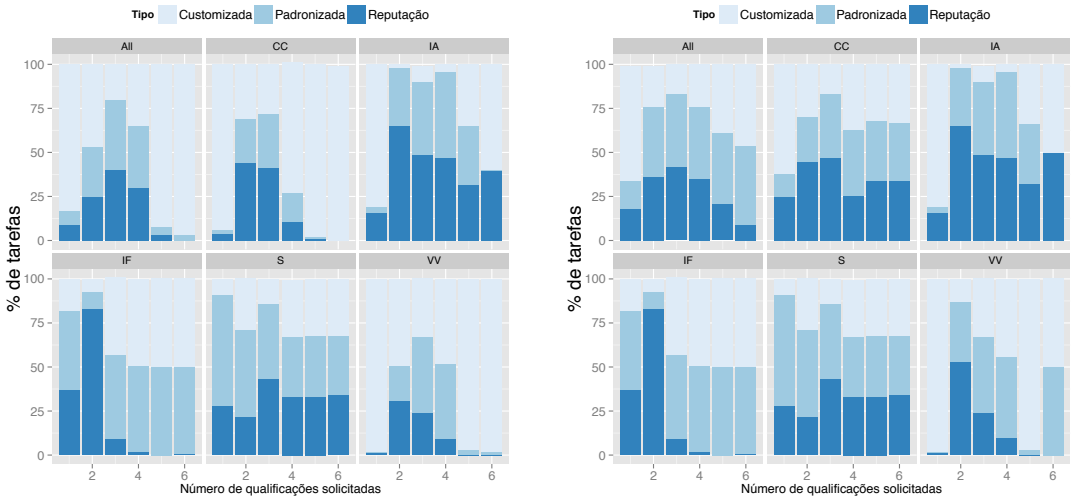
(b) Apenas a cauda da distribuição

Figura 4.9: Distribuição do uso de diferentes tipos de qualificação nas diferentes classes de tarefas do conjunto de dados I.

Ao analisar apenas as tarefas apresentadas pelos solicitantes na cauda da distribuição do conjunto de dados I, verifica-se, novamente, que apenas as classes CC e VV são visivelmente afetadas. No entanto, a ordem de preferência para os diferentes tipos é mantida, com apenas mudanças na porcentagem associada a cada tipo e, naturalmente, os resultados globais também são impactados.

A Figura 4.10 apresenta os dados da distribuição do uso das qualificações para o conjunto de dados II. Quando analisa-se todo o conjunto de dados (4.10a), os resultados mostram que há uma variação considerável na distribuição dos testes de qualificação nas diferentes classes. A classe S apresenta predominância do teste de qualificação do tipo padronizada, principalmente quando uma ou duas qualificações são utilizadas. Quando três qualificações

são usadas, os tipos reputação e padronizada apresentam o mesmo percentual de tarefas. Para quatro, cinco ou seis qualificações, as tarefas usam de forma similar os três tipos de qualificações.



(a) Todas as tarefas

(b) Apenas a cauda da distribuição

Figura 4.10: Distribuição do uso de diferentes tipos de qualificação nas diferentes classes de tarefas do conjunto de dados II.

As tarefas da classe IF utilizam com mais frequência qualificações do tipo padronizada, com exceção quando duas qualificações são usadas (nesse caso o tipo reputação predomina). Percebe-se que na classe IF, quando as tarefas usam de três a seis qualificações, o percentual de tarefas que usa qualificação do tipo padronizada é semelhante ao uso do tipo customizada.

As classes CC e VV possuem distribuições similares, especialmente para os casos de tarefas com um, cinco e seis e tipos de qualificação, em que as qualificações customizadas prevalecem. As tarefas que usam apenas uma única qualificação customizada são mais frequentes nessas classes. O comportamento dessas classes se assemelham ao comportamento do conjunto de dados como um todo. Isso ocorre porque os resultados gerais são fortemente afetados pelo fato de que tarefas de CC são muito mais frequentes do que qualquer outra classe (Figura 4.4).

A classe IA apresenta um comportamento diferente das demais, com a qualificação do tipo customizada prevalecendo nas tarefas que usam uma ou seis qualificações, o tipo reputação nas tarefas com duas ou três qualificações, a combinação padronizada com customizada nas tarefas que usam quatro qualificações e, nas tarefas com cinco qualificações, os três tipos são utilizados.

Quando apenas as tarefas submetidas pelos solicitantes da cauda da distribuição do conjunto de dados II são analisadas (Figura 4.10b), apenas as classes CC e VV são visivelmente afetadas. A classe VV apresenta modificação na ordem de preferência e proporção para as tarefas com duas (o tipo reputação passa a prevalecer seguido do tipo padronizado) e seis qualificações (o tipo customizada diminui em 51%). Na classe CC, as tarefas com uma ou duas qualificações apresentam mudança nas proporções mas mantêm a ordem de preferência dos tipos e, as tarefas com quatro, cinco e seis qualificações apresentam mudanças nas proporções e os três tipos de qualificações são usados de forma parecida. Naturalmente, os resultados globais também são impactados.

Em resumo, os resultados mostram que há uma variação considerável na distribuição de tipos de qualificação nas diferentes classes de tarefas. As tarefas podem variar em relação ao número de qualificações utilizadas e em relação ao tipo de qualificação na mesma classe. Assim, a partir dessas análises, parece que a quantidade e o tipo de qualificação utilizados nas tarefas não são determinados pela classe da tarefa.

4.2.5 Como os solicitantes usam as qualificações

Esta seção tem como objetivo investigar como os testes de qualificação estão associados às tarefas sob a perspectiva dos solicitantes.

Os solicitantes foram agrupados considerando a similaridade entre suas tarefas. Para isso foi utilizada a combinação do algoritmo de agrupamento hierárquico Ward (Ward, 1963) com o algoritmo não-hierárquico k-means (Hartigan and Wong, 1979). Para cada conjunto de dados (I e II) foram gerados sete grupos de solicitantes. Essencialmente, os sete grupos foram formados com base na classe de tarefa que é mais comum no portfólio do solicitante. O

portfólio de um solicitante é definido pela porcentagem de tarefas submetidas pelo solicitante em cada uma das seis classes de tarefas consideradas na etapa de classificação das tarefas. Os valores obtidos para cada classe formam os atributos usados pelo algoritmo de agrupamento.

O portfólio que representa o centro de cada um desses grupos[2], bem como o número de solicitantes em cada grupo, é apresentado na Tabela 4.10 para o conjunto de dados I e, na Tabela 4.11, para o conjunto de dados II.

Tabela 4.10: Descrição dos grupos de solicitantes do conjunto de dados I.

Grupo	**# Solicitantes (%)**	**Portfólio das tarefas**					
		%IF	**%VV**	**%IA**	**%CC**	**%S**	**%N**
G1	1160 (14.5%)	0.01	0.01	0.00	99.96	0.00	0.02
G2	2027 (25.3%)	0.00	0.00	0.00	0.00	100.00	0.0
G3	3179 (39.6%)	0.00	0.00	0.00	0.00	0.00	100.00
G4	631 (7.9%)	0.55	0.43	0.80	4.62	65.50	28.13
G5	126 (1.6%)	11.54	10.29	1.24	15.89	11.03	50.05
G6	248 (3.1%)	0.51	0.14	89.13	1.16	0.67	8.39
G7	648 (8.1%)	99.99	0.00	0.00	0.00	0.00	0.00

Tabela 4.11: Descrição dos grupos de solicitantes do conjunto de dados II.

Grupo	**# Solicitantes (%)**	**Portfólio das tarefas**					
		%IF	**%VV**	**%IA**	**%CC**	**%S**	**%N**
G1	532 (12,8%)	0,02	24,03	0,24	71,79	0,23	3,70
G2	1526 (36,8%)	0,02	0,02	0,38	0,73	97,74	2,23
G3	1191 (28,8%)	0,51	0,17	1,50	2,11	2,25	93,93
G4	295 (7.1%)	0,34	0,34	2,18	9,98	51,09	35,91
G5	218 (5.3%)	0,30	17,72	0,68	46,24	0,11	34,79
G6	262 (6,3%)	0,07	0,11	95,54	0,78	0,67	3,83
G7	118 (2,8%)	79,63	14,21	0,25	0,39	0,10	5,43

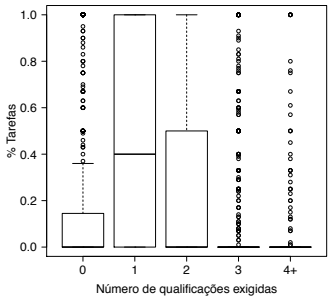
[2]Este é o resultado da execução do algoritmo Ward, que é usado como entrada do algoritmo K-means.

Em cada tabela percebe-se, claramente, para cada grupo de solicitantes, quais tarefas são principalmente submetidas. Por exemplo, os grupos G1, G2, G3, G6 e G7 correspondem aos solicitantes que principalmente enviam tarefas das classes CC, S, N, IA e IF, respectivamente; os grupos G4 e G5, são formados por solicitantes que submetem tarefas em todas as classes, porém no grupo G4, os solicitantes submetem mais tarefas das classes S e N, enquanto que no grupo G5, mais tarefas da classe N e CC.

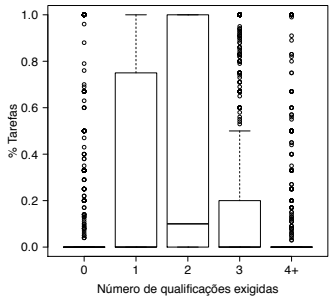
Cada grupo, em cada conjunto de dados, foi analisado com o objetivo de entender como os solicitantes, que têm um portfólio de tarefas semelhante, se comportam em relação à distribuição dos testes de qualificação nas tarefas.

Para o conjunto de dados I, foi analisada, inicialmente, a porcentagem de tarefas com 0, 1, 2, 3 e 4 ou mais testes de qualificação em seu portfólio. A Figura 4.11 apresenta os *box-plots* para cada um dos grupos apresentados na Tabela 4.10 considerando o número de testes de qualificação. Os resultados mostram que, em relação ao número de qualificações utilizadas, um único teste de qualificação é mais comum, sendo claramente mais utilizado pela maioria dos solicitantes nos grupos G1, G3, G5, G6 e G7, e, razoavelmente importante para um número considerável de solicitantes nos grupos G2 e G4. O uso de dois testes de qualificação é claramente mais utilizado pelos solicitantes do grupo G2 e G4, sendo também razoavelmente importante para os solicitantes nos grupos G1, G3, G5, G6 e G7. Um número relativamente pequeno de solicitantes nos grupos G2 e G4 tem uma pequena fração de seu portfólio usando três testes de qualificação, e os grupos G1 e G6 possuem uma quantidade evidente de solicitantes que possuem uma porcentagem considerável de tarefas submetidas sem testes de qualificação.

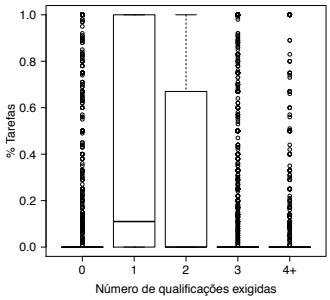
Para o conjunto de dados II foi analisada a porcentagem de tarefas com 0, 1, 2, 3, 4, 5, 6 e 7 ou mais testes de qualificação em seu portfólio. A Figura 4.12 apresenta os *box-plots* para esses atributos para cada um dos grupos apresentados na Tabela 4.11. Os resultados mostram que, em relação ao número de qualificações utilizadas, um teste de qualificação é mais comum, sendo claramente mais utilizado pela maioria dos solicitantes nos grupos G6 e G7 e, razoavelmente importante para um número considerável de solicitantes nos grupos G1, G3, G4 e G5, e, não sendo importante para os solicitantes do grupo G2. O uso de



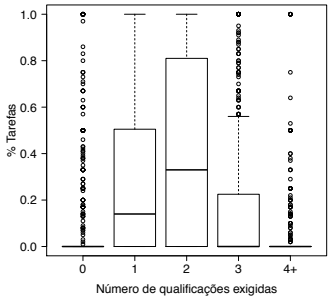
(a) Grupo G1



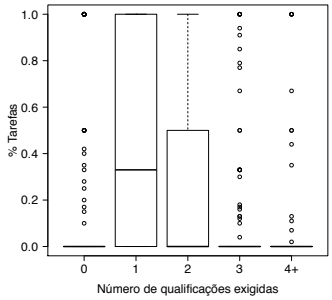
(b) Grupo G2



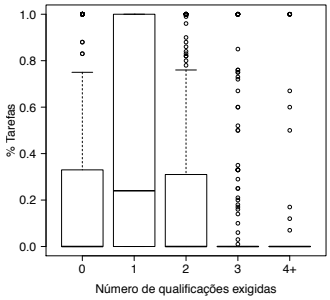
(c) Grupo G3



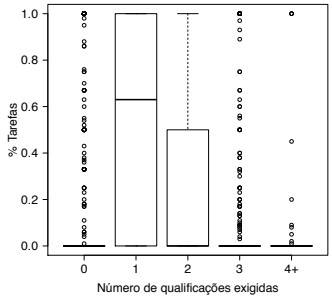
(d) Grupo G4



(e) Grupo G5



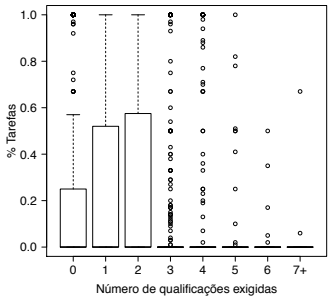
(f) Grupo G6



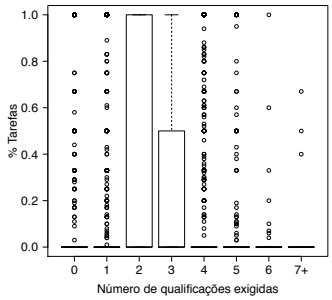
(g) Grupo G7

Figura 4.11: Distribuição do número de qualificações exigidas nos diferentes grupos de solicitantes do conjunto de dados I.

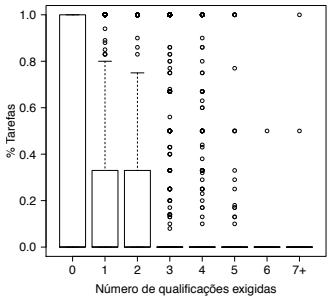
dois testes de qualificação é claramente mais utilizado pelos solicitantes do grupo G2, sendo também razoavelmente importante para os solicitantes nos grupos G1, G3, G4 e G5. Um número relativamente pequeno de solicitantes nos grupos G2 e G4 tem uma pequena fração de seu portfólio usando três testes de qualificação, e somente o G3 possui uma quantidade evidente de solicitantes que possuem uma percentagem notável de tarefas submetidas sem testes de qualificação. Os grupos G1, G5 e G7 também possuem solicitantes com tarefas sem qualificação.



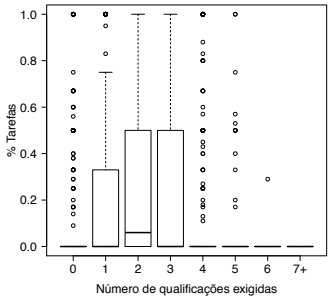
(a) Grupo G1



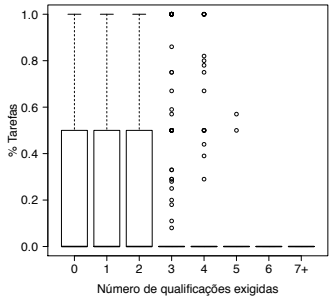
(b) Grupo G2



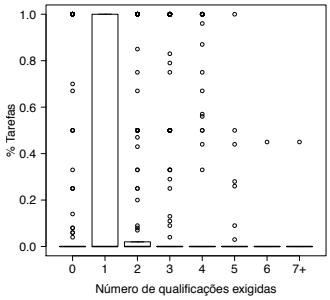
(c) Grupo G3



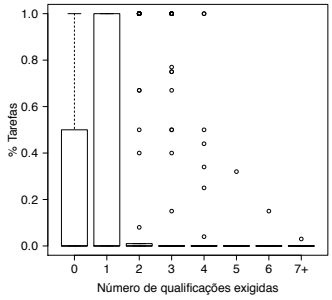
(d) Grupo G4



(e) Grupo G5



(f) Grupo G6



(g) Grupo G7

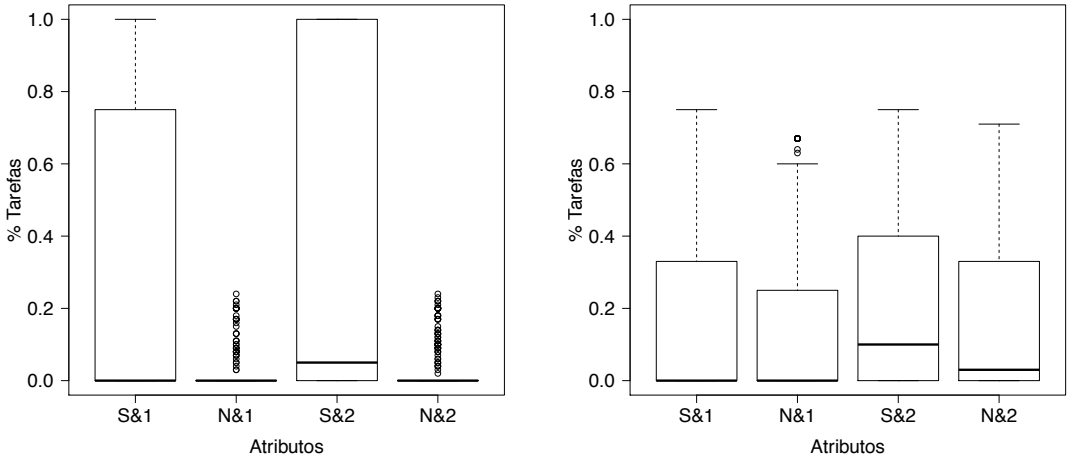
Figura 4.12: Distribuição do número de qualificações exigidas nos diferentes grupos de solicitantes do conjunto de dados II.

Logo, tem-se que, para ambos os conjuntos de dados, nos grupos de solicitantes gerados, o uso de um e dois testes de qualificação são as opções mais comuns de número de testes de qualificação.

No entanto, para solicitantes da classe S, o mais comum é o uso de dois testes de qualificação. Essa conjectura foi investigada analisando a distribuição dos seguintes novos atributos para os solicitantes: i) porcentagem de tarefas da classe S que utilizam 1 teste de qualificação

(S&1); ii) percentual de tarefas da classe N que utilizam 1 teste de qualificação (N&1); iii) percentual de tarefas da classe S que utilizam 2 testes de qualificação (S&2); e, iv) percentagem de tarefas da classe N que utilizam 2 testes de qualificação (N&2).

Os *box-plots* para esses atributos para os grupos G2 e G4 do conjunto de dados I e II são apresentados, respectivamente, na Figura 4.13 e Figura 4.14, pois esses são os grupos que apresentam quantidade considerável de tarefas da classe S, evidenciando de que este é, de fato, o caso.

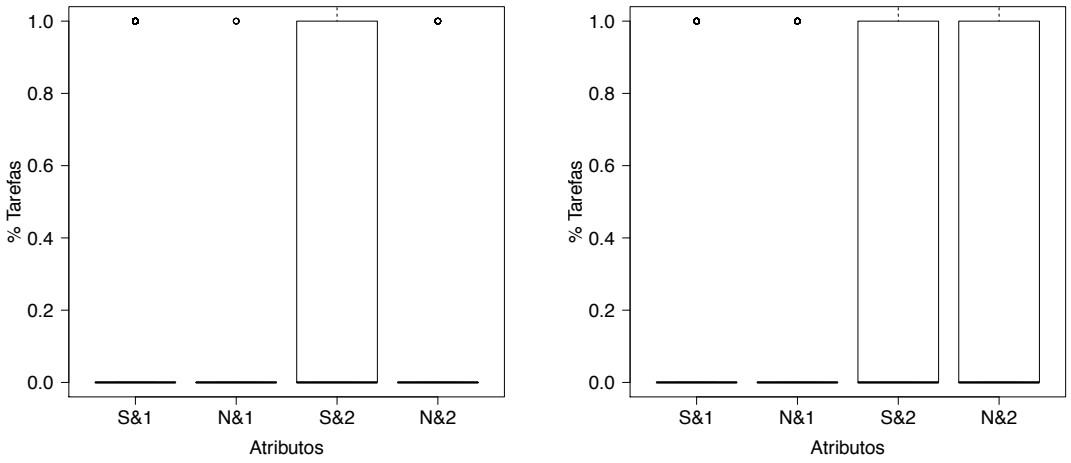


(a) Grupo G2 (b) Grupo G4

Figura 4.13: Distribuição do uso de 1 e 2 qualificações para tarefas das classes S e N considerando os grupos G2 e G4 do conjunto de dados I.

Para a análise do comportamento dos solicitantes em relação à distribuição dos tipos de qualificação nas tarefas, para cada grupo, foram gerados *box-plots* considerando a porcentagem de tarefas para os seguintes testes de qualificação: apenas reputação (R), apenas padronizado (P), apenas customizado (C), apenas a combinação reputação-padronizada (R&P), apenas a combinação reputação-customizada (R&C), apenas a combinação padronizada-customizada (P&C) e, os três tipos juntos (R&P&C).

A Figura 4.15 mostra os resultados para o conjunto de dados I indicando que a qualificação mais utilizada nas classes é o tipo reputação (R), seguida pela combinação do tipo



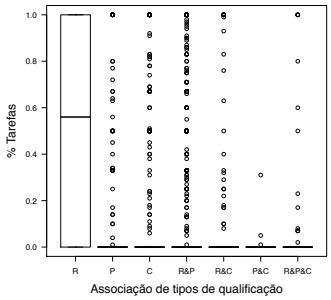
(a) Grupo G2 (b) Grupo G4

Figura 4.14: Distribuição do uso de 1 e 2 qualificações para tarefas das classes S e N considerando os grupos G2 e G4 do conjunto de dados II.

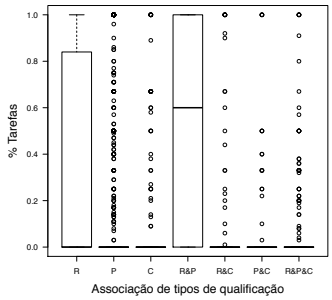
reputação com padronizada (R&P). Nos grupos G1, G6 e G7, a qualificação do tipo reputação (R) é a única que está presente em uma fração substancial de solicitantes, apesar de ser considerável também nos demais grupos. Nos grupos G2 e G4, a combinação do tipo reputação com padronizada (R&P) é a mais comumente utilizada pelos solicitantes, que também é considerável nos grupos G3 e G5.

A Figura 4.16 mostra que, para o conjunto de dados II, o teste de qualificação mais utilizado é a combinação de reputação com padronizada (R&P). Nos grupos G1, G2, G3 e G4, essa combinação é o único tipo de qualificação presente em uma porcentagem considerável de solicitantes, apesar de ser considerável nos demais grupos, excetuando-se o grupo G7. Nos grupos G6 e G7, o teste de qualificação reputação é o mais utilizado pela maioria dos solicitantes, sendo este considerável também nos grupos G1 e G5.

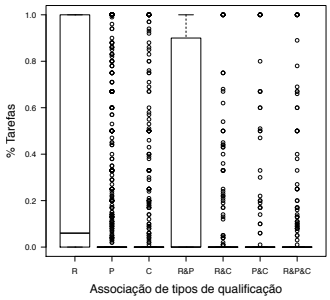
Esses resultados sugerem que, quando os solicitantes usam uma única qualificação, eles tendem a usar o tipo reputação (R), enquanto que quando eles usam duas qualificações, eles tendem a associar uma qualificação do tipo reputação combinada com uma qualificação do tipo padronizada (R&P). O uso de três qualificações ou mais, e a qualificação do tipo customizada (C) aparece apenas em uma porcentagem muito pequena no portfólio da maioria



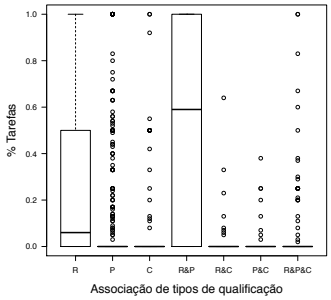
(a) Grupo G1



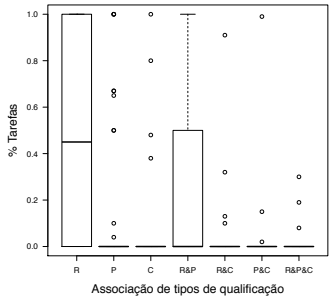
(b) Grupo G2



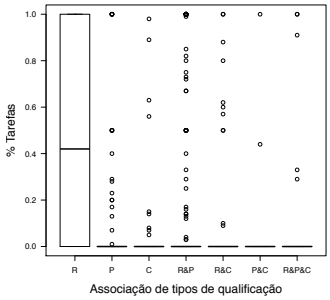
(c) Grupo G3



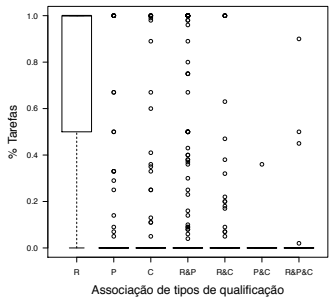
(d) Grupo G4



(e) Grupo G5



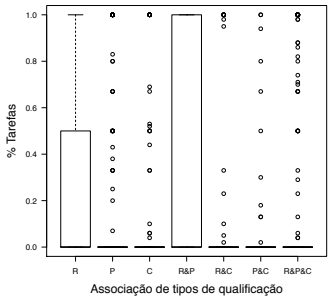
(f) Grupo G6



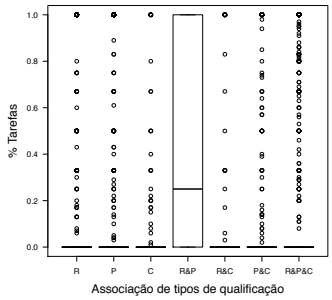
(g) Grupo G7

Figura 4.15: Distribuição do tipo de qualificação exigida nos diferentes grupos de solicitantes do conjunto de dados I.

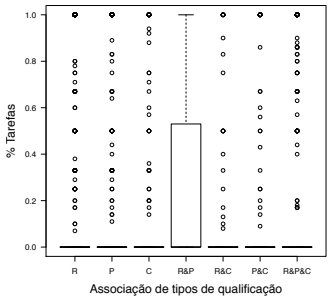
dos solicitantes. Novos atributos foram definidos para investigar esta questão: i) porcentagem de tarefas que exigem uma qualificação do tipo reputação (1&R); ii) porcentagem de tarefas que exigem uma única qualificação padronizada (1&P); iii) porcentagem de tarefas que requerem duas qualificações do tipo reputação (2&R); iv) porcentagem de tarefas que requerem duas qualificações padronizadas (2&P); v) porcentagem de tarefas que requerem uma qualificação do tipo reputação e uma do tipo padronizada (2&R&P). A Figura 4.17 e a Figura 4.18 mostram os *box-plots* para esses atributos para todos os grupos, para o conjunto



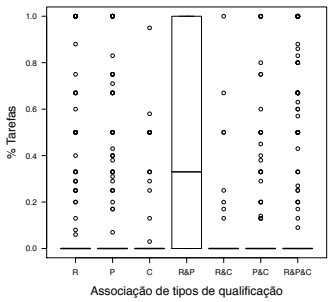
(a) Grupo G1



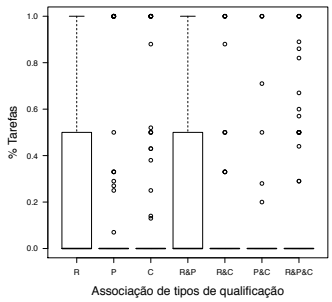
(b) Grupo G2



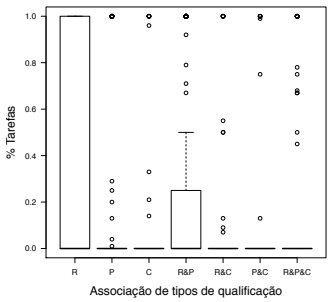
(c) Grupo G3



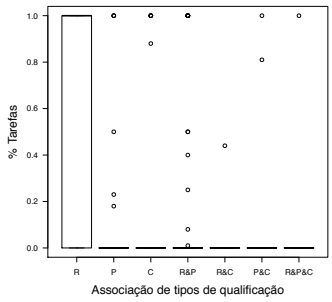
(d) Grupo G4



(e) Grupo G5



(f) Grupo G6

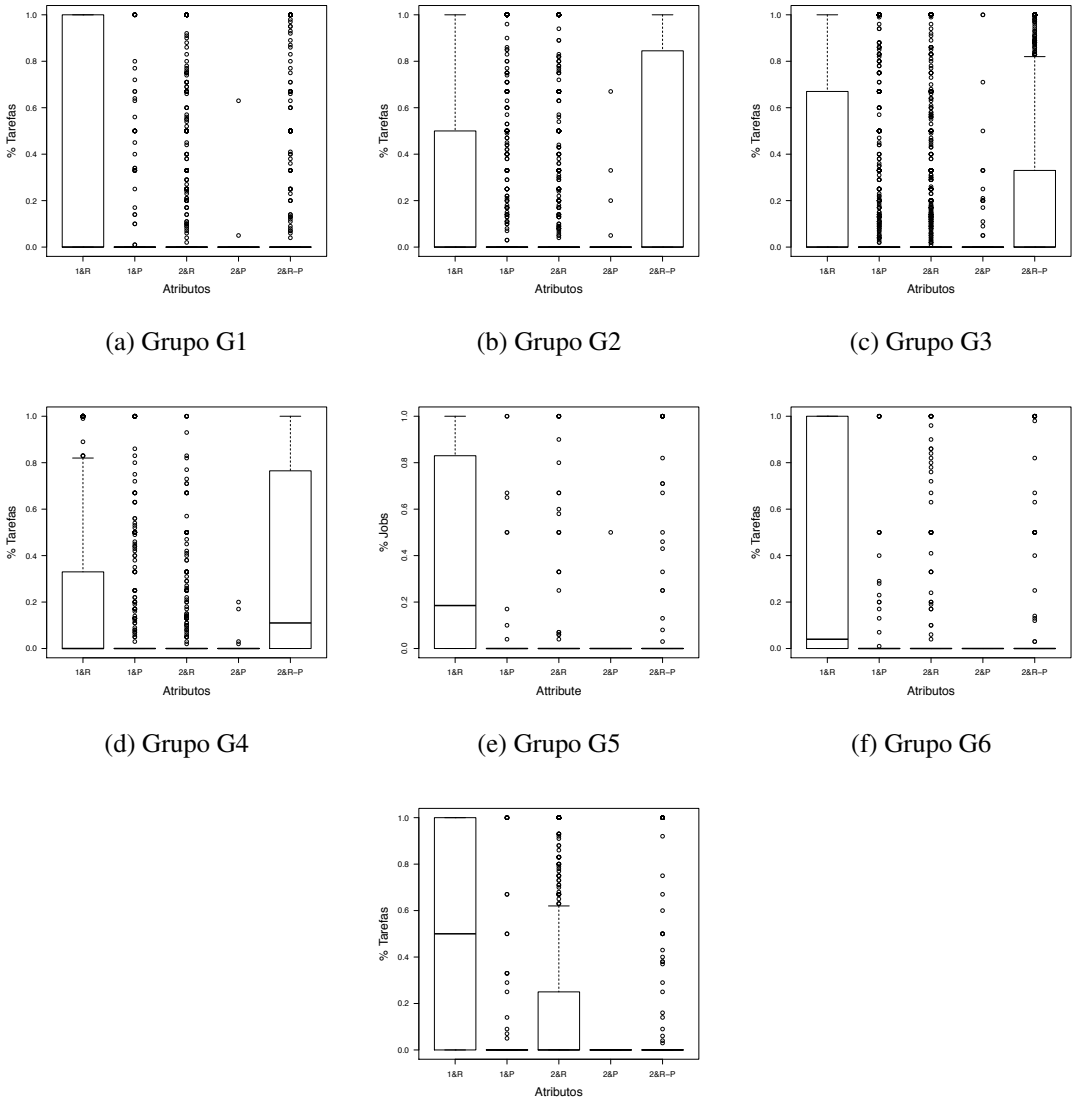


(g) Grupo G7

Figura 4.16: Distribuição do tipo de qualificação exigida nos diferentes grupos de solicitantes do conjunto de dados II.

de dados I e II, respectivamente, o que fornece evidências para suportar essa afirmação.

Em resumo, os resultados mostram que a maioria dos solicitantes utiliza apenas um único teste de qualificação do tipo reputação nas tarefas que submetem, exceto aqueles que enviam tarefas da classe S, que normalmente usam uma qualificação do tipo reputação combinada com uma qualificação do tipo padronizada. Essa combinação também é utilizada em outras classes de tarefas, mas não de forma considerável. Uma das razões para este comportamento



(a) Grupo G1 (b) Grupo G2 (c) Grupo G3

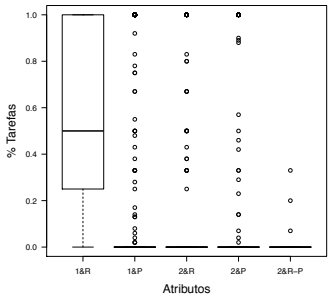
(d) Grupo G4 (e) Grupo G5 (f) Grupo G6

(g) Grupo G7

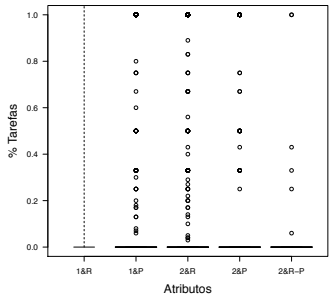
Figura 4.17: Distribuição de 1 e 2 exigências de qualificações do tipo Reputação e Padronizada para cada grupo de solicitantes do conjunto de dados I.

é que esses testes de qualificação estão disponíveis na plataforma facilitando o uso dos mesmos por parte dos solicitantes.

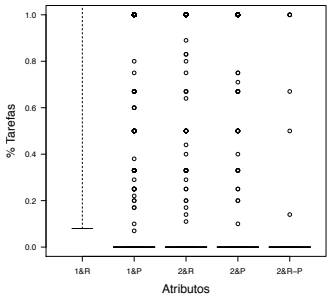
Apesar dos solicitantes R1 e R2 não serem os mesmos nos dois conjuntos de dados analisados, eles apresentam comportamento completamente diferente dos demais solicitantes. No conjunto de dados I, os dois solicitantes mais ativos são os solicitantes *CastingWords* e *Dave*, que juntos submeteram 50,1% das tarefas e, no conjunto de dados II os solicitantes



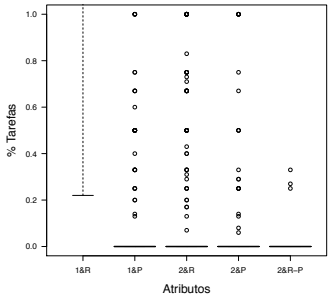
(a) Grupo G1



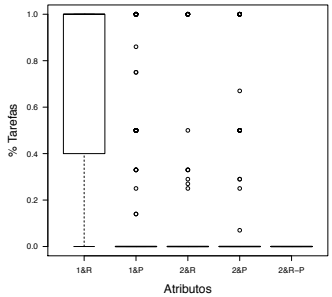
(b) Grupo G2



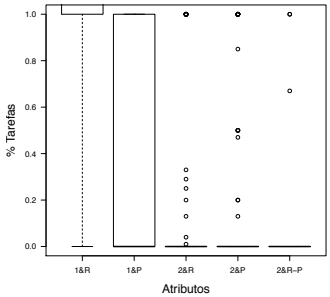
(c) Grupo G3



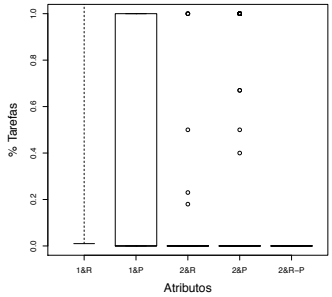
(d) Grupo G4



(e) Grupo G5



(f) Grupo G6



(g) Grupo G7

Figura 4.18: Distribuição de 1 e 2 exigências de qualificações do tipo Reputação e Padronizada para cada grupo de solicitantes do conjunto de dados II.

mais ativos são *Speechpad* e *Ad Tagger*, que juntos submeteram 54% das tarefas.

No conjunto de dados I, o solicitante *CastingWords* é responsável pela submissão de 141.813 tarefas (38,6%) e o solicitante *Dave* por 42.179 tarefas (11,5%). Todas as tarefas submetidas pelos dois usam teste de qualificação do tipo customizado variando apenas a quantidade de testes na tarefa. O solicitante *CastingWords* usa predominantemente um único teste e as tarefas pertencem à classe *Content and Creation* (CC). Já o solicitante *Dave* usa

apenas um teste de qualificação em tarefas da classe *Verification and Validation* (VV) e usa dois testes em tarefas da classe CC, que correspondem, respectivamente, a 31% e 69% do conjunto de tarefas submetidas por este solicitante.

No conjunto de dados II, o solicitante *Speechpad* é responsável por 36,2% de todas as tarefas submetidas no conjunto. Todas as tarefas usam teste de qualificação do tipo customizado sendo que 50,5% de suas tarefas usam 5 testes de qualificações e 34,8% usam 6 testes de qualificação. As tarefas são predominantemente das classes CC e VV. O solicitante *Ad Tagger* submeteu 11.085 tarefas (18% de todo o conjunto), todas usando teste de qualificação. Mais da metade de suas tarefas (59,6%) usa um único teste de qualificação e são da classe CC e, 40,4% das tarefas usam dois testes de qualificação e são da classe *None*.

A participação dos dois solicitante mais ativos de cada conjunto de dados foi investigada em relação ao outro conjunto de dados. O solicitante *Speechpad* não submeteu tarefas no conjunto de dados I. O solicitante *Ad Tagger* submeteu 786 tarefas no conjunto de dados I (0,21% das tarefas submetidas), todas com qualificação do tipo customizada e a maioria (90,6%) usando apenas um único teste de qualificação e sendo da classe *None*.

No conjunto de dados II, o solicitante *Dave* submeteu apenas uma única tarefa, esta sendo da classe CC e usando 3 testes de qualificação do tipo reputação. Essa única tarefa foi submetida no final da coleta de dados, de modo que pode não representar fielmente o comportamento desse solicitante nesse conjunto de dados. O solicitante *CastingWords* submeteu 759 tarefas no conjunto de dados II (1,22% de todas as tarefas), todas usando teste de qualificação do tipo customizado. Quando apenas um único teste de qualificação é usado na tarefa (30% das tarefas), 96% das tarefas são da classe CC. Tarefas submetidas com dois testes de qualificação (68,6%) são também predominantemente da classe CC (96,5%).

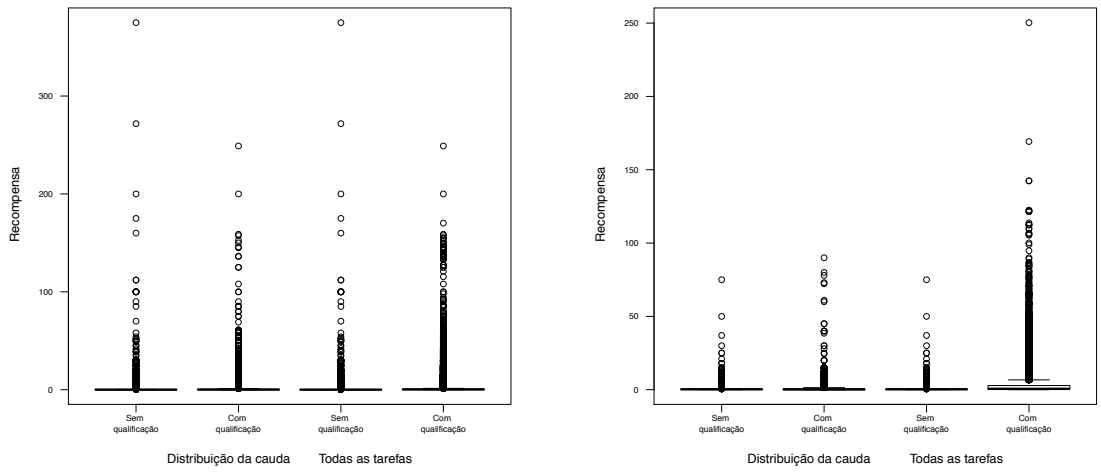
Os dados revelam que os solicitantes *CastingWords* e *Ad Tagger* apresentam comportamento similar nos dois conjuntos de dados em relação ao uso de teste de qualificação, isto é, eles sempre usam, predominantemente, um único teste de qualificação do tipo customizado. O solicitante *CastingWords* parece submeter tarefas sempre do mesmo tipo, isto é, da classe

CC, já o solicitante *Ad Tagger*, submeteu tarefas das classes CC e None, no conjunto de dados II, e, apenas da classe None no conjunto de dados I.

Dessa forma, é possível concluir que os solicitantes mais ativos têm um comportamento pré-definido em relação à plataforma submetendo tarefas usando exclusivamente testes de qualificação customizados, enquanto a grande maioria dos solicitantes, que juntos submetem menos tarefas que os dois mais ativos, utilizam preferencialmente os testes de qualificação disponíveis na plataforma.

4.2.6 Distribuição da recompensa

As tarefas que usam teste de qualificação e tarefas que não usam teste de qualificação foram comparadas em relação à recompensa oferecida (Figura 4.19). Foi realizado teste estatístico para verificar a normalidade da distribuição e verificou-se que os dois grupos apresentam distribuição não normal com assimetria positiva e valores extremos. Os dados apresentados entre parênteses referem-se apenas às tarefas submetidas pelos solicitantes na cauda da distribuição, isto é, sem os dois solicitantes mais ativos (R1 e R2).



(a) Conjunto de dados I

(b) Conjunto de dados II

Figura 4.19: Distribuição da recompensa dos conjuntos de dados I e II.

No conjunto de dados I, os valores médios da recompensa foram US$ 0,97 (US$ 0,91) e US$ 0,65 (US$ 0,51) para tarefas com qualificação e tarefas sem qualificação, respectivamente. No conjunto de dados II, os valores médios da recompensa foram, respectivamente, US$ 3,32 (US$ 0,66) e US$ 0,39 (US$0,39) para as tarefas que utilizam mecanismo de pré-seleção do trabalhador e para as tarefas que não usam. O intervalo de confiança de cada grupo contém a respectiva média, com 95% de confiança, e não inclui o zero. Não há sobreposição dos intervalos de confiança.

Ao examinar as tarefas que não utilizam testes de qualificação, observa-se que não há diferença nos resultados que consideram todo o conjunto de dados e aqueles que consideram apenas parte dele, nos dois conjuntos de dados I e II. Isso ocorre porque quase todas as tarefas enviadas por R1 e R2 pré-selecionam trabalhadores.

Para o conjunto de dados I, a recompensa mais comum é de cinco centavos de dólares (21% das tarefas), 16,2% das tarefas oferecem uma recompensa de dois centavos ou menos, 46,3% oferecem uma recompensa de até cinco centavos, 37,4% oferecem uma recompensa superior ou igual a 10 centavos e 5,1% oferecem uma recompensa superior a um dólar. Para o conjunto de dados II, a recompensa mais comum é de trinta centavos de dólares (45,6% das tarefas), 6,9% das tarefas oferecem uma recompensa de dois centavos ou menos, 20,4% oferecem uma recompensa de até cinco centavos, 77% oferecem uma recompensa superior ou igual a 10 centavos e 4,9% oferecem uma recompensa superior a um dólar.

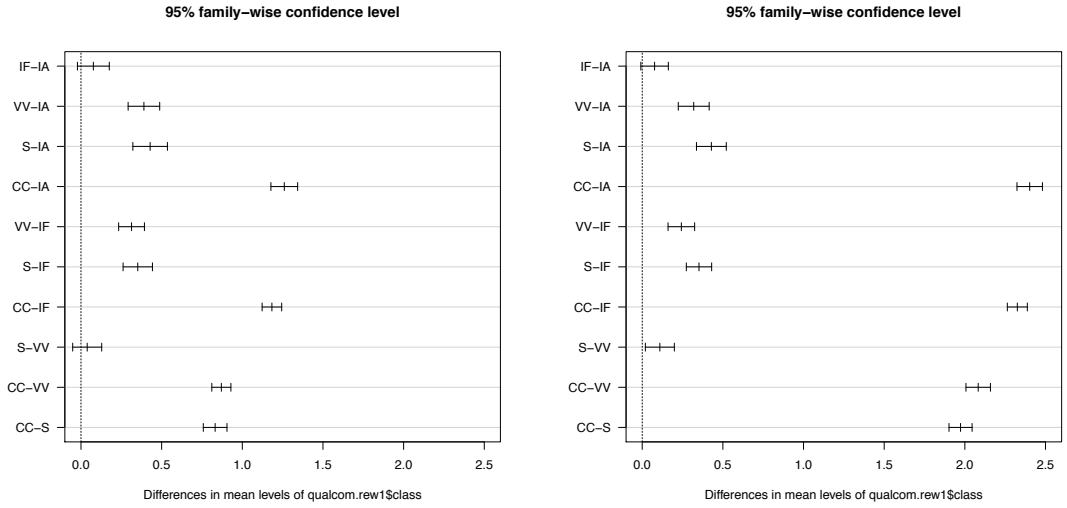
As tarefas que usam teste de qualificação, no conjunto de dados I, apresentam a seguinte distribuição: 6,7% (13,4%) oferecem uma recompensa de dois centavos ou menos, 18,8% (37,0%) oferecem uma recompensa de até cinco centavos, 77,6% (56,9%) oferecem uma recompensa de 10 centavos ou mais e 15,5% (15,5%) oferecem mais de um dólar. No conjunto de dados II, apresentam a seguinte distribuição: 8,5% (20,4%) oferecem uma recompensa de dois centavos ou menos, 14,23% (30,6%) oferecem uma recompensa de até cinco centavos, 84,5% (67,2%) oferecem uma recompensa de 10 centavos ou mais e 33,7% (11,36%) oferecem mais de um dólar.

Foi realizado o teste F para testar a homogeneidade das variâncias cujo resultado mostrou que os dois grupos de tarefas (que usa e que não usa qualificação) apresentam variâncias diferentes, nos dois conjuntos de dados. Para verificar se as diferenças entre as médias dos grupos são grandes o suficiente para concluir que as diferenças ocorrem somente devido à influência da variável independente, quer dizer, se a tarefa exige ou não qualificação do trabalhador, foi realizado o teste *t* não pareado. Com *p-value* <0.05, pode-se afirmar, com 95% de confiança, que existem diferenças significativas entre a recompensa oferecida pelos dois grupos de tarefas. Portanto, as tarefas com teste de qualificação oferecem uma maior recompensa do que as tarefas que não usam teste de qualificação, seja para todo o conjunto de dados seja apenas para a cauda da distribuição.

A recompensa oferecida foi analisada também considerando a classificação de tarefas, a fim de identificar se existem classes que oferecem recompensas melhores do que outras. Apenas as tarefas que usam teste de qualificação foram analisadas. O teste de significância utilizado foi ANOVA (Zar, 2007). O nível de significância considerado em todos os casos foi de 0,05. Verificou-se que existe uma diferença significativa entre a recompensa oferecida pelas cinco classes ao considerar todo o conjunto de dados, bem como quando se considera apenas parte dela. Assim, utilizamos o teste Tukey HSD (Zar, 2007) para determinar quais os pares de classes têm diferenças significativas.

A Figura 4.20 apresenta as diferenças entre as médias dos diferentes pares de classe (nível de confiança de 95%) para o conjunto de dados I. Pares com diferenças significativas são aqueles com limites inferiores positivos. Os pares IF-IA (considerando todas as tarefas ou apenas a cauda da distribuição) e S-VV (Figura 4.20a) são os pares que não apresentam diferenças significativas. Os demais pares contribuíram para as diferenças entre as médias de recompensa detectada pela ANOVA.

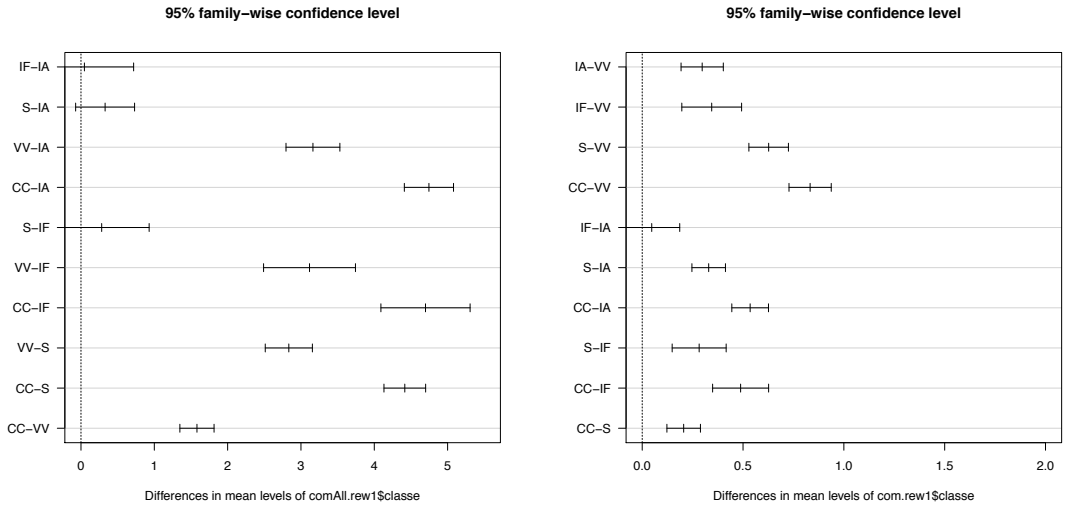
A Figura 4.21 apresenta as diferenças entre as médias dos diferentes pares de classe (nível de confiança de 95%) para o conjunto de dados II. Considerando todo o conjunto de dados (Figura 4.21a) tem-se que os pares IF-IA, S-IA e S-IF não apresentaram diferenças significativas enquanto os demais pares contribuíram para as diferenças entre as médias de recompensa detectada pela ANOVA. Considerando apenas a cauda da distribuição, isto é,



(a) Todas as tarefas (b) Apenas a cauda da distribuição

Figura 4.20: Diferenças entre as médias das recompenses oferecidas pelos pares distintos de classes de tarefas do conjunto de dados I.

sem considerar as tarefas dos dois solicitantes mais ativos (Figura 4.21b), tem-se que apenas o par IF-IA não apresentou diferença significativa enquanto os demais pares contribuíram para as diferenças entre as médias de recompensa detectada pela ANOVA.



(a) Todas as tarefas (b) Apenas a cauda da distribuição

Figura 4.21: Diferenças entre as médias das recompensas oferecidas pelos pares distintos de classes de tarefas do conjunto de dados II.

Verifica-se que as maiores diferenças ocorrem quando a classe CC está presente em ambos os conjuntos de dados. A classe CC tem a média mais alta nos dois conjuntos de dados considerados. Para o conjunto de dados I, tem-se que N = 216.921 e média = US$ 1,32, para todo o conjunto de dados e, N = 46.086 e média = US$ 2,47, quando apenas parte do conjunto de dados é considerado e, para o conjunto de dados II, N = 24470, e média = US$ 5,14, para todo o conjunto de dados e, N = 4037 e média = US$ 0,93, quando apenas parte do conjunto de dados é considerado.

4.3 Considerações finais

Este capítulo apresentou os resultados de um estudo exploratório quantitativo para entender como os testes de qualificação são utilizados na plataforma MTurk sob a perspectiva dos solicitantes. Foram considerados três tipos de testes de qualificação no estudo: reputação, padronizado e customizado. Os testes de qualificação do tipo reputação e padronizado são testes disponíveis na plataforma enquanto que os testes do tipo customizado são aqueles criados pelos próprios solicitantes para atender às suas necessidades particulares.

Verificou-se que a grande maioria das tarefas utiliza algum tipo de teste de qualificação com o objetivo de pré-selecionar trabalhadores na plataforma. Contudo, não existe uma forte relação entre a classe da tarefa e o número e tipo do teste de qualificação que aparecem nessas tarefas. Por outro lado, os resultados mostram que a maioria dos solicitantes tendem a usar apenas um tipo de teste de qualificação dos que existem disponíveis na plataforma, isto é, reputação e padronizado. No entanto, os solicitantes mais ativos, utilizam exclusivamente qualificações do tipo customizada. Isso pode ser uma indicação de que os testes de qualificação existentes na plataforma não sejam suficientes para fornecer o nível de filtragem exigido pelos solicitantes que investem mais pesadamente na plataforma.

As análises também mostram que as tarefas que utilizam teste de qualificação para pré-selecionar o trabalhador oferecem recompensas maiores do que as tarefas que não fazem. Além disso, considerando a classificação de tarefas, existem classes que oferecem recompensas melhores do que outras.

Capítulo 5

O efeito do teste de qualificação customizado na qualidade dos resultados

Muitos sistemas de computação por humanos usam mercados de trabalho *crowdsourcing* de microtarefas para o recrutamento de trabalhadores. O grande desafio é a garantia de qualidade dos resultados submetidos pelos trabalhadores para a solução dos problemas. Algumas soluções são apresentadas na literatura, mas, a habilidade do trabalhador de realizar a tarefa é na maioria das vezes deixada de lado.

Uma alternativa utilizada com o intuito de considerar a habilidade do trabalhador é a pré-seleção do trabalhador através do uso de testes de qualificação. O teste de qualificação é um mecanismo que possibilita ao solicitante avaliar as habilidades do trabalhador para uma determinada tarefa com base em requisitos estabelecidos. Ao utilizar o teste de qualificação o solicitante tem a expectativa de que o trabalhador selecionado produza resultados de qualidade.

Além de delinear como os testes de qualificação são utilizados em sistemas de trabalho online de microtarefas, também é importante conhecer se existe influência desses na qualidade dos resultados.

Avaliar a influência do uso de teste de qualificação, do tipo customizado, na pré-seleção de trabalhadores é importante principalmente para os solicitantes que poderão submeter tarefas

com mais confiança em mercados de trabalho *crowdsourcing* de microtarefas. Além do que, é uma importante contribuição para o problema de recrutamento de trabalhadores adequados em mercados de trabalho *crowdsourcing*.

Este capítulo tem como objetivo, portanto, descrever o estudo realizado para investigar se o uso de testes de qualificação, do tipo customizado, em tarefas submetidas em mercados de trabalho *crowdsourcing* de microtarefas, implica na obtenção de resultados mais precisos.

5.1 Materiais e métodos

Esse estudo é baseado em dados obtidos de experimentos realizados em plataforma de trabalho on-line de microtarefas. Os experimentos e os métodos utilizados na avaliação do efeito do uso de testes de qualificação na qualidade dos resultados obtidos nesse tipo de plataforma são detalhados a seguir.

5.1.1 Descrição dos experimentos

O experimento a ser realizado tem como objetivo avaliar o impacto do uso de teste de qualificação customizado nos resultados produzidos pelos trabalhadores na plataforma MTurk. Cada execução do experimento usa a mesma tarefa em três diferentes cenários: i) todos os trabalhadores podem executar a tarefa; ii) apenas os trabalhadores que possuem uma qualificação específica podem executar a tarefa; iii) somente os trabalhadores mestres podem executar a tarefa. Como explicado anteriormente, os trabalhadores mestres são trabalhadores com alta reputação que são pré-selecionados pela própria plataforma com base em sua performance em trabalhos anteriores.

Ao invés de criar novas tarefas para a execução do experimento, tarefas ativas na plataforma na ocasião do experimento foram escolhidas e utilizadas. A seleção das tarefas utilizou os seguintes critérios:

- A tarefa deveria ser possível de ser realizada por qualquer trabalhador, independente de localização geográfica, faixa etária, sexo, etc.;

- Não deveria ser complexa ou que exigisse demasiado conhecimento prévio para que qualquer trabalhador pudesse realizar;
- Não deveria ser muito fácil, isto é, o trabalhador deveria seguir instruções fornecidas para poder realizar;
- A tarefa deveria possuir um conjunto de respostas a ser utilizado na verificação dos resultados submetidos pelos trabalhadores;
- A tarefa deveria ser do tipo objetiva, isto é, deveria ser possível verificar a resposta através de um conjunto de respostas conhecido. Logo, tarefas que envolvem escrita e tradução de textos, assim como, questionários não se adéquam ao propósito, e,
- A tarefa a ser realizada deveria ser possível de ser realizada por qualquer trabalhador e, obviamente, por trabalhador qualificado (mestre ou pré-selecionado).

Além disso, foi considerado na escolha das tarefas o fato de que as mesmas fossem uma boa representação do conjunto de classes identificadas no Capítulo 4.

Para cada tipo de tarefa, três experimentos diferentes foram executados, cada um contendo 50 HITs da tarefa original, escolhidos de forma aleatória, totalizando, assim, 1350 tarefas.

Os três experimentos diferem entre si em relação ao uso ou não de teste de qualificação, ou se usam trabalhadores mestres.

Os três tipos de experimentos, portanto, são:

- A tarefa pode ser executada por todos os trabalhadores da plataforma. Ou seja, a tarefa não utiliza qualquer teste de qualificação.
- A tarefa é executada apenas por trabalhadores mestres, trabalhadores com alta reputação na plataforma .
- A tarefa é executada apenas por trabalhadores que obtiveram *pontuação* >= *75* no teste de qualificação do tipo customizado utilizado na tarefa.

A comparação do desempenho dos trabalhadores nos três cenários diferentes baseia-se na precisão dos trabalhadores que participaram de cada cenário. A precisão de um trabalhador é definida como a relação entre o número de tarefas, cujos resultados são corretos, e o número total de tarefas para as quais o trabalhador produziu resultados.

5.1.2 Testes de qualificação

Para pré-selecionar os trabalhadores para as tarefas selecionadas para os experimentos, foram criados quatro testes de qualificação customizados (Tabela 5.1) e submetidos na plataforma MTurk. Dessa forma, a habilidade do trabalhador para o experimento pôde ser testada. Ao trabalhador só foi permitido participar do teste de qualificação uma única vez.

Tabela 5.1: Testes de qualificações do tipo habilidade utilizados nos experimentos

Teste de qualificação	**Descrição**	**Experimento**
Find 01- teste de qualificação	Achar e-mail ou forma de contato	Find 01 e Find 02
Find 03- teste de qualificação	Achar nome e endereço de e-mail	Find 03
Image transcription - teste de qualificação	Extrair informação de recibo	Todos os experimentos Image/text Transcription
Categorize - teste de qualificação	Categorizar produto	Todos os experimentos Categorize Image

Os testes de qualificações consistiam de tarefas semelhantes às tarefas selecionadas para os experimentos, com cada tarefa contendo quatro questões. Para cada teste de qualificação foram disponibilizadas 50 tarefas. Para cada trabalhador que aceitou realizar o teste de qualificação, e submeteu os resultados, foi atribuído uma pontuação. A pontuação foi calculada simplesmente dividindo o número de questões corretas do trabalhador por quatro. Assim, os possíveis escores foram: 0, 0,25, 0,5, 0,75 e 1, e cada teste de qualificação tinha como objetivo qualificar até 50 trabalhadores. As tarefas de qualificação ofereceram uma recompensa igual à recompensa oferecida pelas tarefas que exigiam a qualificação.

5.2 Apresentação e análise dos resultados

Com o objetivo de avaliar o desempenho de trabalhadores, foram escolhidos modelos de tarefas que estavam ativos na plataforma MTurk na ocasião da preparação dos experimentos, e, que atendiam aos critérios estabelecidos na metodologia descrita. As tarefas escolhidas

pertencem às seguintes classes: *Information Find* (IF), *Interpretation and Analysis* (IA) e *Content Creation* (CC). Estas são classes com boa representação de tarefas e que permitem avaliar a corretude dos resultados apresentados pelos trabalhadores nos experimentos. As classes das tarefas escolhidas, assim como as subclasses a que pertencem e a ação a ser executada pelo trabalhador em cada classe de tarefa, são apresentadas na Tabela 5.2.

Tabela 5.2: Tabela das ações escolhidas para os experimentos

Classe	Subclasse	Ação
Information Find (IF)	Metadata Finding	Find
Interpretation and Analysis (IA)	Categorization	Categorize Image
Content Creation (CC)	Media Transcription	Image/text transcription

Para cada uma das três classes escolhidas, foram selecionadas três tarefas distintas. A descrição de cada tarefa e de seus atributos (duração da tarefa, recompensa e qualificação exigida) são apresentados na Tabela 5.3. As tarefas submetidas para cada uma das classes escolhidas para a realização dos experimentos são apresentadas no Apêndice A. No experimento foram utilizados os mesmos valores dos atributos das tarefas originais submetidas na plataforma MTurk por solicitantes considerados ativos, com exceção do atributo *qualificação exigida* quando esta foi do tipo customizada. Nesse caso, a qualificação customizada usada na tarefa original foi substituída por um dos testes de qualificação criados nesse trabalho (vide Tabela 5.4). Dessa forma, as tarefas escolhidas apresentam uma diferença em relação às qualificações exigidas. Em relação à recompensa ofertada, tem-se que a média mais baixa é a do grupo de tarefas do tipo *Image/text Transcription* (US$ 0,04). A duração das tarefas apresenta variação entre dez minutos e uma hora.

Os testes de qualificação foram projetados e associados a novos tipos de qualificação criados, com base nos trabalhos selecionados para o experimento. A Tabela 5.4 fornece uma descrição dos testes de qualificação projetados. Foram criados dois tipos distintos de teste de qualificação para as tarefas da subclasse *Metadata Finding* devido às diferentes habilidades necessárias para executar diferentes tarefas nesta subclasse. Para as outras subclasses, um único tipo de qualificação por subclasse foi suficiente para qualificar os trabalhadores para

Tabela 5.3: Descrição das tarefas utilizadas nos experimentos.

Experimento	**Descrição da tarefa**	**Subclasse**	**Duração (min.)**	**Recompensa (US$)**	**Qualificação original exigida**
Categorize-1	Categorizar imagem	Categorization	60	0.02	Customizada
Categorize-2	Indicar se uma categoria associada a um produto está correta ou não	Categorization	10	0.06	Customizada
Categorize-3	Dada uma lista de categorias, escolher a categoria correta para um produto	Categorization	10	0.06	Customizada
Find-1	Achar endereço de e-mail de universidades americanas	Metadata Finding	30	0.04	Master
Find-2	Identificar companhias ou organizações de URL's fornecida	Metadata Finding	60	0.05	Master
Find-3	Achar nomes e endereço de e-mail de autor de artigo	Metadata Finding	60	0.05	Master
Transcribe-1	Classificar recibo	Media Transcription	20	0.02	Nenhuma
Transcribe-2	Escrever as palavras apresentadas na imagem	Media Transcription	15	0.05	Nenhuma
Transcribe-3	Extrair informação de recibo de compra	Media Transcription	60	0.05	Nenhuma

Tabela 5.4: Tipos de qualificações customizadas criadas para os experimentos.

Qualificação	**Descrição**	**Experimento**
Categorize-Qualification	Categorizar produtos	Todos os experimentos para a subclasse categorization
Find-1&2-Qualification	Achar endereço de e-mail ou forma de contato	Find-1 and Find-2
Find-3-Qualification	Achar nome e endereço de e-mail	Find-3
Transcribe-Qualification	Extrair informação de recibos de compras	todos os experimentos para a subclasse media transcription

os três tipos de tarefas diferentes usadas no experimento para essas subclasses.

A distribuição da pontuação obtida pelos trabalhadores que executaram as tarefas de cada um dos quatro tipos de testes de qualificação criados é apresentada na Figura 5.1. A qualifi-

cação *Transcribe-Qualification* foi a que obteve o maior número de trabalhadores (34 trabalhadores ou 68%) com a pontuação mínima necessária para a seleção (0,75). A qualificação *Categorize-Qualification*, por outro lado, foi a que obteve o menor número de trabalhadores com a pontuação mínima exigida (20 trabalhadores ou 40%). Para a qualificação *Find-1&2*, 28 trabalhadores atingiram uma pontuação igual ou superior a 0,75 (56%) e, 26 trabalhadores (52%) conseguiram para a qualificação *Find-3*.

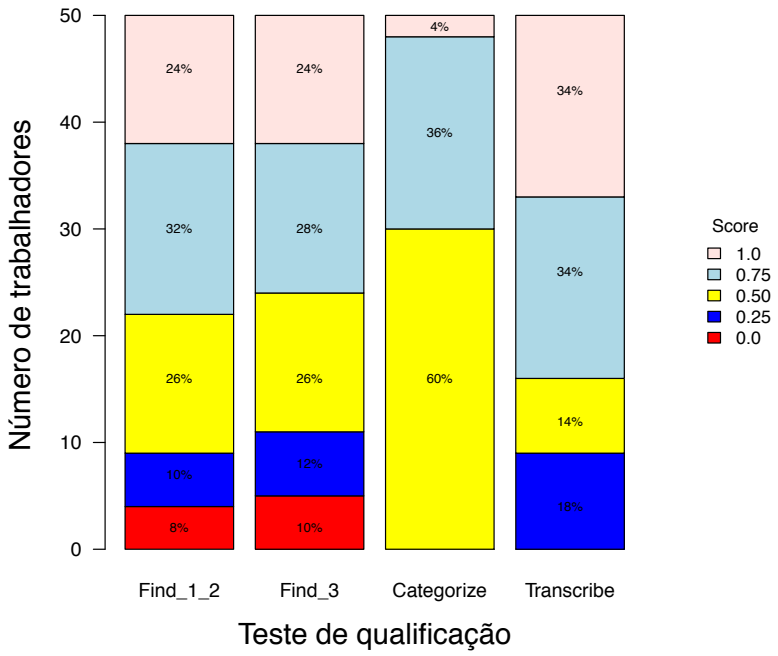


Figura 5.1: Distribuição da pontuação para os testes de qualificação.

Considerando todos os trabalhadores que executaram as tarefas de qualificação, a pontuação média obtida foi de 0,71 para *Transcribe-Qualification*, 0,64 para *Find-1&2-Qualification* e 0,61 para as qualificações *Categorize-Qualification* e *Find-3-Qualification*, enquanto as medianas foram, respectivamente, 0,75, 0,75, 0,50 e 0,75. Todos os trabalhadores que executaram a qualificação *Categorize-Qualification* conseguiram atingir uma pontuação de pelo menos 0,50, enquanto que para a qualificação *Transcribe-Qualification*, a pontuação mínima foi de 0,25. As qualificações *Find-1&2* e *Find-3* foram as únicas duas qualificações em que alguns trabalhadores apresentaram uma pontuação de 0,00.

Considerando apenas os trabalhadores que se qualificaram, ou seja, alcançaram uma pontuação igual ou superior a 0,75, a pontuação média foi de 0,88 para *Transcribe-Qualification*, 0,86 para *Find-1&2-Qualification* e *Find-3-Qualification*, e 0,78 para *Categorize-Qualification*, enquanto as medianas foram, respectivamente, 0,88, 0,75, 0,75 e 0,75.

Após a execução dos testes de qualificação para pré-selecionar os trabalhadores, o experimento foi submetido para avaliar o impacto da pré-seleção de trabalhadores sobre a precisão dos resultados alcançados. Para cada experimento, três novas tarefas foram criadas, de acordo com a tarefa original selecionada, obtendo assim um total de 27 execuções (3 classes de tarefas / 3 tarefas por classe / 3 critérios de seleção de trabalhadores). Não foi possível evitar que um trabalhador pré-selecionado ou um trabalhador mestre também realizasse tarefas que fossem disponibilizadas para qualquer trabalhador na plataforma. No entanto, verificações de consistência foram realizadas, a posteriori, para identificar essas situações e remover os resultados correspondentes. Foi realizada apenas a remoção de um resultado de um trabalhador mestre e que forneceu uma resposta tanto para o experimento que exigia a qualificação de mestre quanto para o experimento que não exigia qualquer qualificação. Em outro caso, um trabalhador mestre fez o trabalho duas vezes, isto é, participou do experimento que exigia trabalhadores mestres e do experimento que exigia a qualificação customizada correspondente. No entanto, uma vez que o trabalhador realizou o teste de qualificação customizado, os dois resultados foram mantidos.

A distribuição da pontuação alcançada pelos trabalhadores que realizaram as diferentes tarefas utilizadas nas execuções do experimento é apresentada na Figura 5.2. Para cada trabalhador que participou de cada um dos experimentos realizados foi calculada a acurácia do seu trabalho considerando o número de respostas corretas em relação ao número total de respostas submetidas pelo trabalhador. Isso foi feito para cada tipo de experimento e de trabalhador. Assim sendo, os pontos do gráfico da Figura 5.2 representam a acurácia de cada trabalhador que participou de cada um dos experimentos.

As tarefas dos experimentos foram projetadas da maneira mais convencional, onde não há limite no número de HITs que um trabalhador tem direito de executar. Assim, o número

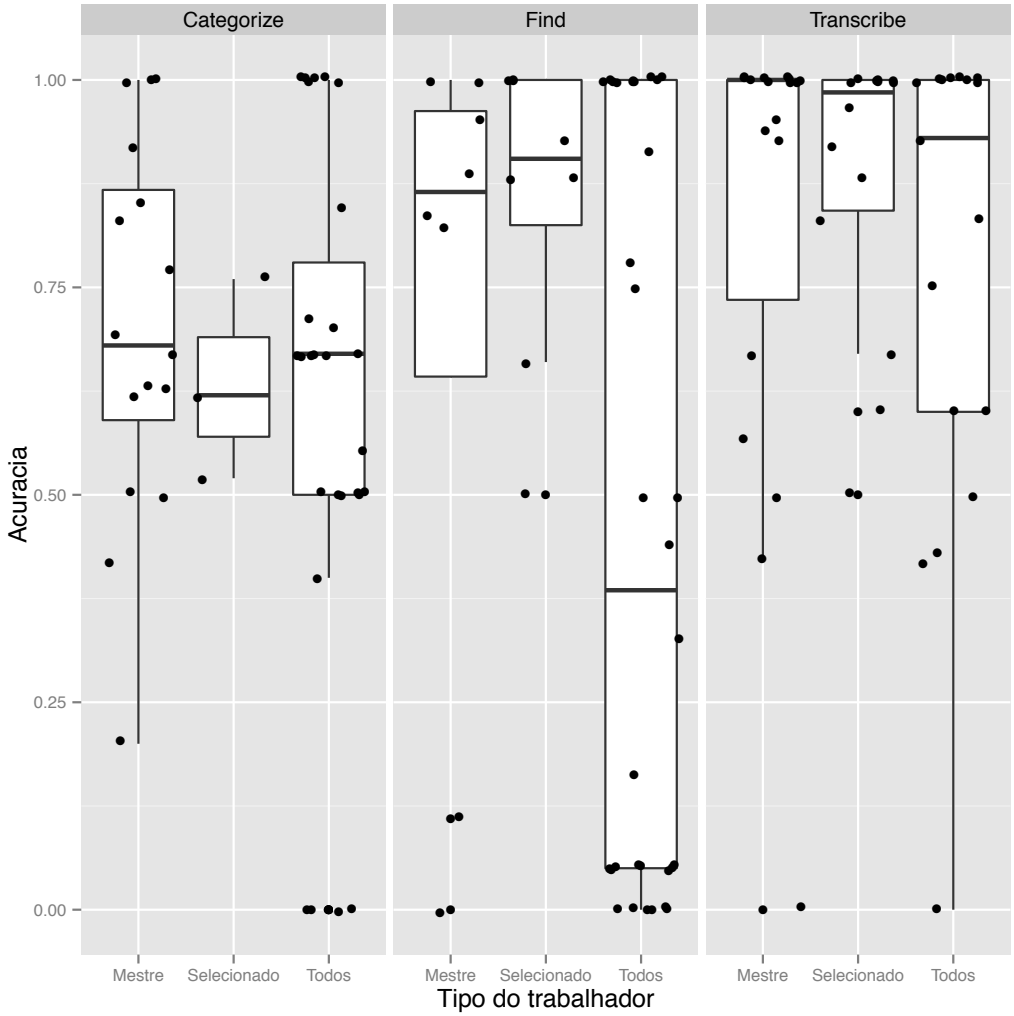


Figura 5.2: Distribuição da pontuação dos trabalhadores nos experimentos.

de trabalhadores participantes em cada execução de uma tarefa variou consideravelmente. Para o experimento *Categorize*, 28 trabalhadores participaram das tarefas que não exigiam qualquer tipo de qualificação, 16 das tarefas que utilizavam os trabalhadores mestres e 3 das tarefas que exigiam trabalhadores pré-selecionados. Para os outros experimentos, esses números foram, respectivamente, 32, 8 e 8 para o experimento *Find*, e 17, 18 e 14 para o experimento *Transcribe*.

Os resultados mostram que, em todos os experimentos, a pontuação média dos trabalhadores pré-selecionados (0,63 para *Categorize*, 0,86 para *Find* e 0,88 para *Transcribe*) é melhor do que a dos trabalhadores que não foram pré-selecionados (respectivamente, 0,60, 0,46 e 0,77). Considerando a mediana, os trabalhadores pré-selecionados apresentam melhor de-

sempenho para os experimentos *Find* (0.90 contra 0.39) e *Transcribe* (0.99 contra 0.93), e pior para o experimento *Categorize* (0.62 contra 0.67). Ressalta-se que, no experimento de *Categorize*, dos 20 trabalhadores que foram qualificados para executar as tarefas no trabalho, apenas 3 realmente as executaram.

Quando comparado com a pontuação média dos trabalhadores mestres (respectivamente, 0,70, 0,70 e 0,83 para os experimentos *Categorize*, *Find* e *Transcribe)*, os trabalhadores pré-selecionados apresentam um resultado mais baixo no experimento *Categorize* e melhor nos experimentos *Find* e *Transcribe*. Considerando a mediana, a comparação é quase a mesma, com a diferença de que os trabalhadores pré-selecionados apresentam resultados um pouco mais baixos do que os trabalhadores mestres no experimento *Transcribe* (0,99 contra 1,00).

O desempenho, ligeiramente pior, apresentado pelos trabalhadores pré-selecionados nos experimentos *Find* e *Categorize*, quando comparado aos trabalhadores mestres, pode estar relacionado às características das classes a que pertencem (IF e IA, respectivamente). Ressalta-se que, ao considerar os dados analisados nos capítulos anteriores, tem-se que quando as tarefas dessas classes usam um único teste de qualificação, o tipo de qualificação mais utilizado é o tipo reputação, isto é, o que está disponível na plataforma. Por outro lado, as tarefas na classe IA, mais frequentemente, usam duas qualificações, com predominância da qualificação baseada em reputação.

A partir das médias amostrais para a acurácia dos trabalhadores em cada experimento realizado, buscou-se obter informações para a população através da determinação de intervalos de confiança usando a distribuição T Student. Assim tem-se que, os intervalos de confiança apresentados na Tabela 5.5 tem probabilidade 0,95 (95% de confiança) de conter o real valor da média da população, considerando cada tipo de tarefa. Assim, é possível inferir que para os experimentos *Find* e *Transcribe* os solicitantes ao usarem trabalhadores pré-selecionados terão resultados melhores, enquanto que para o experimento *Categorize*, os mestres podem apresentar resultados melhores.

Como esperado, a pontuação alcançada dos trabalhadores pré-selecionados pode ser aumentada se o solicitante for mais restritivo ao estabelecer a pontuação mínima para qualificar

Tabela 5.5: Intervalo de confiança para a média da acurácia da população

Experimento	Trabalhador	Mínimo	Máximo	Erro
Categorize	Mestre	0,58	0,82	0,12
	Selecionado	0,33	0,93	0,30
	Todos	0,48	0,73	0,13
Find	Mestre	0,36	1,04	0,34
	Selecionado	0,70	1,01	0,15
	Todos	0,30	0,62	0,16
Transcribe	Mestre	0,69	1,04	0,14
	Selecionado	0,78	0,98	0,10
	Todos	0,62	0,92	0,15

os trabalhadores. Se apenas os trabalhadores pré-selecionados que obtiveram pontuação com valor igual a 1 (pontuação máxima) forem considerados, as pontuações médias para os experimentos *Find* e *Transcribe* aumentam para 0.88 e 0.93, respectivamente[1].

Analisando o tempo de execução de cada experimento tem-se que o experimento que finalizou em menor tempo (menos de 48 horas nos três cenários considerados) foi o *Categorize* levando a crer que este tipo de tarefa atrai mais rapidamente os trabalhadores, porém, os resultados apresentados não foram os melhores. Fato que corrobora com a questão de que muitos trabalhadores procuram aumentar seus rendimentos executando tarefas no menor tempo possível. Já o experimento *Transcribe*, para os trabalhadores mestres e pré-selecionados, apresentou tempo de execução de 31 e 35 dias, respectivamente, e melhores resultados. O mesmo ocorreu no experimento *Find* para os trabalhadores pré-selecionados, isto é, o experimentou demorou 12 dias para ser finalizado, mas também com bons resultados. A exceção ficou conta do experimento *Find* para trabalhadores mestres que demorou 46 dias para ser finalizado, levando a crer que tarefas desse tipo atraem pouco os trabalhadores mestres.

[1]Quando se atribui o valor 1 para a pontuação mínima necessária para a seleção, nenhum dos três trabalhadores que executaram as tarefas do experimento *Categorize* teriam sido qualificados, portanto, não é possível calcular um novo valor de pontuação média para esse caso.

5.3 Considerações finais

Neste capítulo foram realizados experimentos para analisar a qualidade dos dados submetidos em plataformas de trabalho on-line de microtarefas por três tipos de trabalhadores: trabalhadores selecionados através de teste de qualificação customizado, que testa a habilidade do trabalhador, trabalhadores mestres e por qualquer trabalhador registrado na plataforma.

A principal contribuição desse capítulo é o estudo detalhado da qualidade dos resultados obtidos de tarefas submetidas na plataforma MTurk usando conjuntos de dados reais. Várias tarefas foram submetidas à plataforma MTurk. Essas tarefas foram criadas a partir de tarefas existentes na plataforma no momento em que o experimento foi executado. Para cada tarefa submetida, os resultados dos três tipos de trabalhadores foram considerados.

Os resultados mostraram que a pontuação média alcançada pelos trabalhadores pré-selecionados foi sempre maior que a alcançada por trabalhadores que não foram pré-selecionados. Além disso, o desempenho de trabalhadores pré-selecionados foi muito próximo dos trabalhadores considerados mestres e, em alguns cenários, ainda melhor.

Portanto, o estudo apresentado neste capítulo mostra que o uso de testes de qualificação do tipo customizado, para pré-selecionar trabalhadores, é satisfatório para determinados tipos de tarefas, em plataforma de trabalho on-line de microtarefas.

Capítulo 6

Conclusões

Esta pesquisa teve como objeto de estudo o uso de mecanismo para pré-selecionar trabalhadores em mercados de trabalho on-line de microtarefas. Investigar a tese de que o uso de teste de qualificação, do tipo customizado, em tarefas submetidas em mercados de trabalho on-line de microtarefas tem influência na qualidade dos resultados foi o objetivo geral. Neste capítulo são apresentados os principais resultados e contribuições. Por fim, são discutidas as limitações existentes neste trabalho, bem como direções para trabalhos futuros.

6.1 Resultados e contribuições

A primeira contribuição desse trabalho foi a análise da plataforma de mercado de trabalho on-line de microtarefas com o objetivo de entender como os solicitantes utilizam estratégias de controle de qualidade de resultados em suas tarefas. Os resultados dessa análise mostraram que o tipo de estratégia mais utilizada é a pré-seleção dos trabalhadores através de teste de qualificação do tipo reputação para executar uma determinada tarefa. No entanto, os solicitantes mais ativos, utilizam exclusivamente qualificações do tipo customizada.

O teste de qualificação do tipo reputação considera o histórico do comportamento do trabalhador na plataforma e está disponível para o solicitante na interface gráfica de criação do projeto de tarefa. Dado que a maioria dos solicitantes submetem poucas tarefas, entende-se que essa maioria usa a plataforma uma única vez ou usa de forma esporádica e, por isso, usar um teste de qualificação do tipo reputação é mais fácil, além do que, o solicitante não tem

um conhecimento mais aprofundado da plataforma.

Os poucos solicitantes dizem respeito aos solicitantes mais regulares na plataforma e portanto também mais experientes na plataforma. Esses solicitantes possuem um comportamento pré-definido em relação à plataforma, submetendo tarefas usando teste de qualificação do tipo customizado, o que pode ser uma indicação de que os testes de qualificação do tipo reputação e padronizado não sejam suficientes para obtenção de resultados no nível desejado por esses solicitantes.

O fato é que, independente do solicitante ser ou não ativo na plataforma, ele usa em suas tarefas algum teste de qualificação (reputação, padronizado ou customizado) para filtrar trabalhadores para suas tarefas, o que corrobora com o fato de que a qualidade dos resultados submetidos pelos trabalhadores é um problema em plataformas online de microtarefas.

A segunda contribuição consistiu em agrupar as tarefas em classes com o intuito de analisar o uso da estratégia de pré-seleção considerando as características das tarefas. Essa análise permitiu verificar que não existe uma forte relação entre a classe da tarefa e o número e tipo do teste de qualificação que aparecem nessas tarefas. Verificou-se também que algumas classes de tarefas remuneram melhor que outras e que estas são as classes que contêm tarefas que usam testes de qualificação do tipo customizado. Uma conclusão disso é que os solicitantes mais ativos usam testes de qualificação do tipo customizado para obter melhores resultados e remuneram melhor seus trabalhadores. Apesar de trabalhos existentes na literatura terem verificado que apenas o aumento de incentivo financeiro não garante melhores resultados, ele pode ser utilizado como uma motivação para o trabalhador que tem seu conhecimento testado anteriormente para poder participar da tarefa. Além disso, imagina-se que os usuários mais ativos possuem também uma reputação considerável na plataforma, atraindo mais trabalhadores para suas tarefas.

A terceira contribuição consistiu em verificar a influência do uso de teste de qualificação, do tipo customizado, na qualidade dos resultados submetidos pelos trabalhadores. A métrica utilizada para a verificação da influência foi a precisão das respostas submetidas em três tipos de tarefas selecionadas: categorizar produtos, busca de informação e trans-

crição de informação. Os resultados obtidos comprovam a validade da hipótese de que os trabalhadores pré-selecionados apresentam melhores resultados quando comparados com os trabalhadores em geral em todos os experimentos. Os resultados mostraram que a pontuação média alcançada pelos trabalhadores pré-selecionados foi sempre maior que a alcançada por trabalhadores que não foram pré-selecionados, nos três tipos de tarefas utilizadas nos experimentos.

Os trabalhadores especialistas da plataforma, denominados de mestres, também foram utilizados nos experimentos com o objetivo de comparar seus resultados com os resultados dos trabalhadores que são pré-selecionados. Ao comparar o desempenho dos trabalhadores pré-selecionados com o desempenho dos trabalhadores considerados mestres no sistema, os trabalhadores pré-selecionados apresentaram melhor desempenho nos experimentos em que é possível usar o teste de qualificação para treinar o trabalhador e cujo treinamento não depende exclusivamente da interpretação do trabalhador. Isso foi possível nos experimentos Find e Transcribe. Logo, a hipótese de que os resultados de trabalhadores pré-selecionados são diferentes dos resultados de trabalhadores mestres é confirmada, e, em alguns casos, melhores.

O solicitante para usar os trabalhadores especialistas em suas tarefas, isto é, os trabalhadores mestres, ele paga à plataforma uma taxa adicional de 5% sobre o custo da tarefa, além da taxa paga à plataforma. O solicitante para usar o teste de qualificação do tipo customizado, em geral, deve submeter na plataforma uma amostra de tarefas semelhantes ao que deseja submeter a posteriori, para fazer a pré-seleção dos trabalhadores. Para atrair trabalhadores para essas tarefas, ele remunera o trabalhador, tendo assim um custo adicional que é calculado considerando o número de trabalhadores que deseja qualificar, a recompensa a ser oferecida e o número de tarefas que o trabalhador deve executar no teste. Logo, quando vale a pena usar trabalhadores mestres e quando vale a pena usar teste do tipo customizado? Será que o custo do uso de trabalhadores mestres influencia a escolha do uso de teste do tipo de qualificação customizada por parte dos solicitantes mais ativos?

Considerando t o número de tarefas, r a recompensa para cada tarefa, c o custo de qualificar um trabalhador para responder essa tarefa e w o número de trabalhadores que se deseja

usar na tarefa, então, para o custo de usar trabalhador mestre ou qualificação customizada ser o mesmo, tem-se que:

$$t.w.r.1,05 = t.w.r + c \tag{6.1}$$

Logo,

$$t = 20.c/(w.r) \tag{6.2}$$

Se *t* for maior ou igual a expressão, então sugere-se ao solicitante usar o teste de qualificação do tipo customizado, caso contrário, é mais viável usar o trabalhador mestre. Dado que a maioria dos solicitantes submete poucas tarefas, o uso de trabalhadores mestres torna-se mais viável. Já para os solicitantes mais ativos, o uso de trabalhadores mestres é inviável financeiramente, além do que, o uso do teste de qualificação do tipo customizado oferece um maior controle por parte dos solicitantes.

Ao realizar os experimentos o que se esperava era que em todos eles os trabalhadores pré-selecionados apresentassem melhores resultados. No entanto, ao analisar os resultados, foi possível concluir que as tarefas de categorização atraem mais rapidamente os trabalhadores e talvez isso aconteça porque esse tipo de tarefa exija, aparentemente, menos tempo do trabalhador. Nessa linha de raciocínio o que se tem é que os trabalhadores veem nesse tipo de tarefa uma possibilidade de aumentar seus rendimentos na plataforma, executando mais tarefas em menos tempo. Essa conclusão foi possível pelo fato do experimento contendo tarefas de categorização ter sido concluído em menos de 48 horas. No entanto, tarefas de categorização são tarefas que exigem interpretação e análise e, portanto, não são tão fáceis. Daí porque os melhores resultados submetidos são os dos trabalhadores considerados mestres na tarefa.

Por outro lado, os experimentos contendo tarefas de transcrição de conteúdo e busca de informação apresentaram comportamento diferente em relação ao tempo do experimento. Estes demoraram mais de 30 dias para serem finalizados. No entanto, esses experimentos eram mais específicos, não exigiam apenas interpretação e análise, e, o trabalhador apto conhecia mais especificamente a tarefa. Logo, os resultados submetidos pelos trabalhadores pré-selecionados foram melhores, independente da reputação do solicitante.

Observa-se, no entanto, que os solicitantes que submetem tarefas específicas usando teste de qualificação do tipo customizado não conseguem resultados num curto espaço de tempo em função do tamanho do conjunto de trabalhadores aptos. O uso de teste de qualificação do tipo customizado demanda tempo tanto do solicitante quanto do trabalhador. O solicitante precisa projetar os testes em função de suas necessidades, disponibilizar na plataforma, selecionar os trabalhadores, e só depois publicar as tarefas apenas para os trabalhadores aptos. O trabalhador, por sua vez, vai usar um tempo para realizar o teste, tempo este que poderia ser usado para realizar outras tarefas, além do que não existe a certeza de que conseguirá êxito no teste de qualificação. Soma-se a isso o fato de que o trabalhador já gasta um certo tempo na busca de tarefas para realizar. Por isso, uma possibilidade para solicitantes que desejam respostas imediatas é não usar testes de qualificação customizados, porém, os resultados podem ser de baixa qualidade.

A quarta contribuição é que esse estudo é um dos primeiros que tenta prover informação sobre o uso de testes de qualificação e seus resultados são valiosos para orientar os solicitantes em melhor definirem suas tarefas e conseguirem resultados de melhor qualidade.

6.2 Limitações

Uma limitação deste estudo é a definição de uma única plataforma de trabalho on-line de microtarefas. Os resultados obtidos nesta pesquisa são dependentes da situação particular desta pesquisa. Logo, os resultados obtidos neste estudo podem ser considerados uma evidência, mas não uma garantia de que os mesmos comportamentos serão observados em outras tarefas ou em outras plataformas. Por isso, apesar da plataforma escolhida ser a de maior popularidade, pesquisa futura deve abranger outras plataformas semelhantes.

Em relação à classificação das tarefas, inicialmente, foi realizada uma classificação manual de uma amostra analisando os atributos título e descrição de cada tarefa. Um segundo avaliador realizou a classificação de uma amostra menor. Como resultado tem-se que é possível que avaliadores diferentes classifiquem a mesma tarefa de forma diferente em função da interpretação individual nos casos em que a tarefa pode ser de fato inserida em mais de uma classe. Os resultados mostraram que a discordância provocaria um possível aumento

nas classes IF e/ou IA e diminuiria o percentual da classe CC, no entanto, a predominância da classe CC continuaria. Sendo assim, pesquisa futura poderia verificar a possibilidade de usar outros métodos para a classificação.

A limitação em relação à métrica utilizada nos experimentos está relacionada ao tipo de tarefa considerado. Foram consideradas apenas tarefas consideradas objetivas, isto é, tarefas cujas respostas podem ser avaliadas quanto à corretude através de uma comparação com um conjunto de respostas conhecidas. Estas tarefas possuem respostas em formato estruturado. Dessa forma, tarefas com respostas em formato não estruturado, como por exemplo, uma redação de texto, são mais difíceis de serem avaliadas diretamente com essa métrica.

6.3 Trabalhos futuros

A pesquisa sobre o efeito do uso de testes de qualificação, do tipo customizado, não foi completamente esgotada neste documento. Portanto, alguns trabalhos que podem evoluir a partir deste, sempre com o objetivo de obter melhores resultados dos trabalhadores, são:

- Reproduzir o estudo em outras plataformas de trabalho on-line de microtarefas e verificar se o comportamento da pontuação dos trabalhadores se repete.
- Incluir mais tipos de tarefas no estudo de avaliação dos testes de qualificação.
- Analisar o efeito do uso de teste de qualificação, do tipo customizado, em conjunto com o teste de qualificação do tipo reputação, na pré-seleção de trabalhadores em mercados de trabalho on-line de microtarefas.
- Verificar se o desempenho dos trabalhadores pré-selecionados com teste de qualificação do tipo customizado sofre influência de esquemas de incentivos financeiros.
- Investigar a correlação entre as motivações do trabalhador para requisitar uma qualificação do tipo customizada e a reputação do solicitante, já que verificou-se que existem testes de qualificação customizados utilizados em muitas tarefas, mas sem trabalhadores qualificados, bem como, testes de qualificação utilizados em poucas tarefas mas com muitos trabalhadores qualificados.

Bibliografia

Agrawal, Rakesh and Ramakrishnan Srikant (1994). Fast algorithms for mining association rules in Large Databases. In Jorge B. Bocca, Matthias Jarke; and Carlo Zaniolo (eds.): *VLDB'94. Proceedings of the* 20th *International Conference on Very Large Data Bases, Santiago, Chile, 12 September – 15 September 1994*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 487–499.

Ahn, Luis Von (2005). Human computation. Tese de Doutorado. Carnegie Mellon University, 2005. UMU Order Number: AAI3205378.

Allahbakhsh, Mohammad; Boualem Benatallah; Aleksandar Ignjatovic; Hamid Reza Motahari-Nezhad; Elisa Bertino; and Schahram Dustdar (2013). Quality Control in Crowdsourcing Systems: Issues and Directions. *IEEE Internet Computing*, vol. 17, no. 2, pp. 76–81.

Alonso, Omar and Stefano Mizzaro (2009). Can we get rid of trec assessors? using mechanical turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, Volume 15. pp. 16.

Amabile, Teresa M; Karl G Hill; Beth A Hennessey; and Elizabeth M Tighe (1994). The work preference inventory: assessing intrinsic and extrinsic motivational orientations. *Journal of personality and social psychology*, vol. 66, no. 5, pp. 950.

Amazon Web Services (2017). Amazon Mechanical Turk - Requester UI Guide. AWS Documentation. Amazon Web Services (AWS), 2017. https://docs.aws.amazon.com/AWSMechTurk/latest/RequesterUI/amt-ui.pdf. Accessed 28 January 2017.

Archak, Nikolay (2010). Money, Glory and Cheap Talk: Analyzing Strategic Behavior of Contestants in Simultaneous Crowdsourcing Contests on TopCoder.Com. In Michael Rappa, Paul Jones, Juliana Freire; and Soumen Chakrabarti (eds.): *WWW'10. Proceedings of the* 19th *International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010*. New York, NY, USA: ACM, pp. 21–30.

Barowy, Daniel W.; Charlie Curtsinger; Emery D. Berger; and Andrew McGregor (2012). AutoMan: A Platform for Integrating Human-based and Digital Computation. In Gary T. Leavens and Matthew B. Dwyer (eds.): *OOPSLA'12. Proceedings of the ACM International Conference on Object Oriented Programming Systems Languages and Applications, Tucson, AZ, USA, 21–25 October 2012*. New York, NY, USA: ACM, pp. 639–654.

Bernstein, Michael S.; Greg Little; Robert C. Miller; Björn Hartmann; Mark S. Ackerman; David R. Karger; David Crowell; and Katrina Panovich (2015). Soylent: a Word Processor with a Crowd Inside. *Communications of the ACM*, vol. 58, no. 8, pp. 85–94.

Bigham, Jeffrey P; Chandrika Jayant; Hanjie Ji; Greg Little; Andrew Miller; Robert C Miller; Robin Miller; Aubrey Tatarowicz; Brandyn White; Samual White; et al. (2010). Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, pp. 333–342.

Brabham, Daren C (2008). Moving the crowd at istockphoto: The composition of the crowd and motivations for participation in a crowdsourcing application. *First monday*, vol. 13, no. 6.

Bueno, M. (2002). As teorias de motivação humana e sua contribuição para a empresa humanizada: um tributo a Abraham Maslow. Revista do Centro de Ensino Superior de Catalão (CESUC), v.4, n. 6, 2002.

Callison-Burch, Chris (2009). Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In Philipp Koehn and Rada Mihalcea (eds.): *EMNLP'09. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–7 August 2009*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 286–295.

Carletta, Jean (1996). Assessing Agreement on Classification Tasks: the kappa Statistic. *Computational Linguistics*, vol. 22, no. 2, pp. 249–254.

Cooper, Seth; Firas Khatib; Adrien Treuille; Janos Barbero; Jeehyung Lee; Michael Beenen; Andrew Leaver-Fay; David Baker; Zoran Popović; et al. (2010). Predicting protein structures with a multiplayer online game. *Nature*, vol. 466, no. 7307, pp. 756–760.

Corney, Jonathan R; Carmen Torres-Sanchez; A Prasanna Jagadeesan; Xiu T Yan; William C Regli; and Hugo Medellin (2010). Putting the crowd to work in a knowledge-based factory. *Advanced Engineering Informatics*, vol. 24, no. 3, pp. 243–250.

Deci, Edward L and Richard M Ryan (2013). Intrinsic motivation and self-determination in human behavior. 1985. *Consultado en septiembre*.

Demartini, Gianluca; Djellel Eddine Difallah; and Philippe Cudré-Mauroux (2012). Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st international conference on World Wide Web*. ACM, pp. 469–478.

Difallah, Djellel Eddine; Gianluca Demartini; and Philippe Cudré-Mauroux (2012). Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms. In Ricardo A. Baeza-Yates, Stefano Ceri, Piero Fraternali; and Fausto Giunchiglia (eds.): *CrowdSearch'12. Proceedings of the First International Workshop on Crowdsourcing Web Search, Lyon, France, 17 April 2012*. Aachen, Germany: CEUR Workshop Proceedings, pp. 26–30.

DiPalantino, Dominic and Milan Vojnovic (2009). Crowdsourcing and all-pay auctions. In *Proceedings of the 10th ACM conference on Electronic commerce*. ACM, pp. 119–128.

Dow, Steven; Anand Kulkarni; Brie Bunge; Truc Nguyen; Scott Klemmer; and Björn Hartmann (2011). Shepherding the Crowd: Managing and Providing Feedback to Crowd Workers. In Desney Tan, Geraldine Fitzpatrick, Carl Gutwin, Bo Begole; and Wendy A. Kellogg (eds.): *CHI EA'11. CHI'11 Extended Abstracts on Human Factors in Computing Systems, Vancouver, BC, Canada, 7–12 May 2011*. New York, NY, USA: ACM, pp. 1669–1674.

Eickhoff, Carsten and Arjen P. de Vries (2011). How Crowdsourcable is your Task? In Matthew Lease, Vitor R. Carvalho; and Emine Yilmaz (eds.): *CSE'10. Proceedings of the SIGIR 2011 Workshop on Crowdsourcing for Search Evaluation, Hong Kong, China, 9 February 2011*. New York, NY, USA: ACM, pp. 11–14.

Estellés-Arolas, Enrique and Fernando González-Ladrón-de Guevara. Towards an integrated crowdsourcing definition. *Journal of Information science, volume=38, number=2, pages=189–200, year=2012, publisher=Sage Publications Sage UK: London, England*.

Feldman, Ronen and James Sanger (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge, UK: Cambridge University Press.

Fischer, Debra A; Megan E Schwamb; Kevin Schawinski; Chris Lintott; John Brewer; Matt Giguere; Stuart Lynn; Michael Parrish; Thibault Sartori; Robert Simpson; et al. (2012). Planet hunters: the first two planet candidates identified by the public using the kepler public archive data. *Monthly Notices of the Royal Astronomical Society*, vol. 419, no. 4, pp. 2900–2911.

Frei, Brent (2009). Paid crowdsourcing: Current state & progress towards mainstream business use. smartsheet white paper.

Gadiraju, Ujwal; Ricardo Kawase; and Stefan Dietze (2014). A Taxonomy of Microtasks on the Web. In Leo Ferres, Gustavo Rossi, Virgilio Almeida; and Eelco Herder (eds.): HT'14. *Proceedings of the* 25th *ACM Conference on Hypertext and Social Media, Santiago, Chile, 1–4 September 2014*. New York, NY, USA: ACM, pp. 218–223.

Grier, David Alan (2013). *When Computers Were Human*. Princeton, NJ, USA: Princeton University Press.

Hartigan, John A. and Manchek A. Wong (1979). Algorithm AS 136: A k-means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108.

Heckhausen, Heinz (1991). Motivation and action (pk leppman, trans.). *Beflin. Heidel*.

Hirth, Matthias; Tobias Hoßfeld; and Phuoc Tran-Gia (2011). Anatomy of a crowdsourcing platform-using the example of microworkers. com. In *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference on*. IEEE, pp. 322–329.

Hirth, Matthias; Tobias Hoßfeld; and Phuoc Tran-Gia (2013). Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. *Mathematical and Computer Modelling*, vol. 57, no. 11, pp. 2918–2932.

Howe, Jeff (2006). The rise of crowdsourcing. *Wired magazine*, vol. 14, no. 6, pp. 1–4.

Hsu, Chih-Wei; Chih-Chung Chang; and Chih-Jen Lin (2016). A Practical Guide to Support Vector Classification. Online guide. National Taiwan University, 2016. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf. Accessed 28 January 2017.

Hu, Chang; Benjamin B Bederson; and Philip Resnik (2010). Translation by iterative collaboration between monolingual users. In *Proceedings of Graphics Interface 2010*. Canadian Information Processing Society, pp. 39–46.

Ipeirotis, Panos (2010a). Be a top mechanical turk worker: You need $5 and 5 minutes. *Blog: Behind Enemy Lines*.

Ipeirotis, Panagiotis G. (2010b). Analyzing the Amazon Mechanical Turk Marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, vol. 17, no. 2, pp. 16–21.

Ipeirotis, Panagiotis G (2010c). Demographics of mechanical turk.

Ipeirotis, Panagiotis G.; Foster Provost; and Jing Wang (2010). Quality Management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation, Washington, DC, USA, 25 July 2010*. New York, NY, USA: ACM, pp. 64–67.

Kaufmann, Nicolas; Thimo Schulze; and Daniel Veit (2011). More than fun and money. worker motivation in crowdsourcing-a study on mechanical turk. In *AMCIS*, Volume 11. pp. 1–11.

Khanna, Shashank; Aishwarya Ratan; James Davis; and William Thies (2010). Evaluating and Improving the Usability of Mechanical Turk for Low-income Workers in India. In Andrew Dearden, Tapan Parikh; and Lakshminarayanan Subramanian (eds.): *ACM DEV '10. Proceedings of the First ACM Symposium on Computing for Development, London, UK, 17–18 December 2010*. New York, NY, USA: ACM, pp. 12:1–12:10.

Khazankin, Roman; Harald Psaier; Daniel Schall; and Schahram Dustdar (2011). QoS-Based Task Scheduling in Crowdsourcing Environments. In Gerti Kappel, Zakaria Maamar; and Hamid R. Motahari-Nezhad (eds.): *ICSOC'11. Proceedings of the* 9th *International Conference on Service-Oriented Computing, Paphos, Cyprus, 5–8 December 2011*. Berlin, Heidelberg: Springer-Verlag, pp. 297–311.

Kittur, Aniket; Ed H. Chi; and Bongwon Suh (2008). Crowdsourcing User Studies with Mechanical Turk. In Mary Czerwinski, Arnie Lund; and Desney Tan (eds.): *CHI'08. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy, 5–10 April 2008*. New York, NY, USA: ACM, pp. 453–456.

Kittur, Aniket; Jeffrey V. Nickerson; Michael Bernstein; Elizabeth Gerber; Aaron Shaw; John Zimmerman; Matt Lease; and John Horton (2013). The Future of Crowd Work. In Amy Bruckman, Scott Counts, Cliff Lampe; and Loren Terveen (eds.): *CSCW'13. Proceedings of the 2013 Conference on Computer Supported Cooperative Work, San Antonio, Texas, USA, 23–27 February 2013*. New York, NY, USA: ACM, pp. 1301–1318.

Kohavi, Ron (1995). A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the* 14th *International Joint Conference on Artificial Intelligence - Vol. 2, Montreal, Quebec, Canada, 20–25 August 1995*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 1137–1143.

Kokkodis, Marios and Panagiotis G. Ipeirotis (2013). Have You Done Anything Like That?: Predicting Performance Using Inter-category Reputation. In Stefano Leonardi, Alessandro Panconesi, Paolo Ferragina; and Aristides Gionis (eds.): *WSDM'13. Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, Rome, Italy, 4–8 February 2013*. New York, NY, USA: ACM, pp. 435–444.

Kosorukoff, Alex (2001). Human based genetic algorithm. In *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, Volume 5. IEEE, pp. 3464–3469.

Kulkarni, Anand; Matthew Can; and Björn Hartmann (2012). Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the acm 2012 conference on computer supported cooperative work*. ACM, pp. 1003–1012.

Lakhani, Karim R; Lars Bo Jeppesen; Peter Andreas Lohse; Jill A Panetta; et al. (2007). *The value of openess in scientific problem solving*. Division of Research, Harvard Business School.

Law, Edith (2011). Defining (human) computation. In *CHI 2011: Workshop on Crowdsourcing and Human Computation*.

Law, Edith and Luis Von Ahn (2011). Human computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 5, no. 3, pp. 1–121.

Law, Edith LM; Luis Von Ahn; Roger B Dannenberg; and Mike Crawford (2007). Tagatune: A game for music and sound annotation. In *ISMIR*, Volume 3. pp. 2.

Le, John; Andy Edmonds; Vaughn Hester; and Lukas Biewald (2010). Ensuring Quality in Crowdsourced Search Relevance Evaluation: The Effects of Training Question Distribution. In Vitor R. Carvalho, Matthew Lease; and Emine Yilmaz (eds.): *CSE'10. Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation, Geneva, Switzerland, 23 July 2010*. New York, NY, USA: ACM, pp. 21–26.

LEVY, Pierre (1999). A inteligência coletiva: por uma antropologia do ciberespaço. São Paulo: Edições Loyola: 1999. Cibercultura. São Paulo: Editora, v. 34, 1999.

Lintott, Chris J; Kevin Schawinski; Anže Slosar; Kate Land; Steven Bamford; Daniel Thomas; M Jordan Raddick; Robert C Nichol; Alex Szalay; Dan Andreescu; et al. (2008). Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, vol. 389, no. 3, pp. 1179–1189.

Little, Greg; Lydia B. Chilton; Max Goldman; and Robert C. Miller (2010). Turkit: human computation algorithms on mechanical turk. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, pp. 57–66.

Malone, Thomas W; Robert Laubacher; and Chrysanthos Dellarocas (2009). Harnessing crowds: Mapping the genome of collective intelligence.

Mao, Andrew; David C Parkes; Ariel D Procaccia; and Haoqi Zhang (2011). Human computation and multiagent systems: an algorithmic perspective. In *Proceedings of the twenty-fifth AAAI conference on artificial intelligence*.

Mason, Winter and Duncan J. Watts (2010). Financial Incentives and the Performance of Crowds. *ACM SigKDD Explorations Newsletter*, vol. 11, no. 2, pp. 100–108.

Oleson, David; Alexander Sorokin; Greg Laughlin; Vaughn Hester; John Le; and Lukas Biewald (2011). Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. In Luis von Ahn and Panagiotis G. Ipeirotis (eds.): *AAAIWS'11. Proceedings of the* 11th *AAAI Conference on Human Computation, San Francisco, CA, USA, 8 August 2011*. Palo Alto, CA, USA: AAAI Press, pp. 43–48.

Ponciano, Lesandro; Francisco Brasileiro; Nazareno Andrade; and Lívia Sampaio (2014). Considering Human Aspects on Strategies for Designing and Managing Distributed Human Computation. *Journal of Internet Services and Applications*, vol. 5, no. 1, pp. 1–15.

Ponciano, Lesandro; Francisco Brasileiro; Robert Simpson; and Arfon Smith (2014). Volunteers' engagement in human computation for astronomy projects. *Computing in Science & Engineering*, vol. 16, no. 6, pp. 52–59.

Quinn, Alexander J. and Benjamin B. Bederson (2011). Human Computation: A Survey and Taxonomy of a Growing Field. In Desney Tan, Geraldine Fitzpatrick, Carl Gutwin, Bo Begole; and Wendy A. Kellogg (eds.): *CHI'11. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, Canada, 7–12 May 2011*. New York, NY, USA: ACM, pp. 1403–1412.

Raykar, Vikas C; Shipeng Yu; Linda H Zhao; Gerardo Hermosillo Valadez; Charles Florin; Luca Bogoni; and Linda Moy (2010). Learning from crowds. *Journal of Machine Learning Research*, vol. 11, no. Apr, pp. 1297–1322.

Rogstadius, Jakob; Vassilis Kostakos; Aniket Kittur; Boris Smus; Jim Laredo; and Maja Vukovic (2011). An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. *ICWSM*, vol. 11, pp. 17–21.

Ross, Joel; Lilly Irani; M Silberman; Andrew Zaldivar; and Bill Tomlinson (2010). Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*. ACM, pp. 2863–2872.

Rossi, Rafael Geraldeli (2011). Representação de coleções de documentos textuais por meio de regras de associação. Tese de Doutorado. Universidade de São Paulo.

Ryan, Richard M and Edward L Deci (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, vol. 25, no. 1, pp. 54–67.

Rzeszotarski, Jeffrey M. and Aniket Kittur (2011). Instrumenting the Crowd: Using Implicit Behavioral Measures to Predict Task Performance. In Jeff Pierce, Maneesh Agrawala; and Scott Klemmer (eds.): *UIST'11. Proceedings of the* 24th *Annual ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, USA, 16–19 October 2011*. New York, NY, USA: ACM, pp. 13–22.

Sauermann, Henry and Chiara Franzoni (2015). Crowd science user contribution patterns and their implications. *Proceedings of the National Academy of Sciences*, vol. 112, no. 3, pp. 679–684.

Schulze, Thimo; Dennis Nordheimer; and Martin Schader (2013). Worker Perception of Quality Assurance Mechanisms in Crowdsourcing and Human Computation Markets. In *Proceedings of the* 19th *Americas Conference on Information Systems, Chicago, IL, USA, 15–17 August 2013*. Red Hook, NY, USA: Curran Associates, Inc., pp. 4046–4056.

Schulze, Thimo; Stefan Seedorf; David Geiger; Nicolas Kaufmann; and Martin Schader

(2011). Exploring task properties in crowdsourcing-an empirical study on mechanical turk. In *ECIS*, Volume 11. pp. 1–1.

Shaw, Aaron D; John J Horton; and Daniel L Chen (2011). Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. ACM, pp. 275–284.

Sheng, Victor S; Foster Provost; and Panagiotis G Ipeirotis (2008). Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 614–622.

Singh, Push; Thomas Lin; Erik Mueller; Grace Lim; Travell Perkins; and Wan Li Zhu (2002). Open mind common sense: Knowledge acquisition from the general public. *On the move to meaningful internet systems 2002: CoopIS, DOA, and ODBASE*, pp. 1223–1237.

Snow, Rion; Brendan O'Connor; Daniel Jurafsky; and Andrew Y. Ng (2008). Cheap and Fast—but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks. In Mirella Lapata and Hwee Tou Ng (eds.): *EMNLP'08. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii, 25–27 October 2008*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 254–263.

Sodré, Ianna and Francisco Brasileiro (2017). An analysis of the use of qualifications on the amazon mechanical turk online labor market. *Computer Supported Cooperative Work (CSCW)*, vol. 26, no. 4-6, pp. 837–872.

Sorokin, Alexander and David Forsyth (2008). Utility data annotation with amazon mechanical turk. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*. IEEE, pp. 1–8.

Su, Qi; Dmitry Pavlov; Jyh-Herng Chow; and Wendell C. Baker (2007). Internet-scale Collection of Human-reviewed Data. In Carey Williamson, Mary Ellen Zurko, Peter Patel-Schneider; and Prashant Shenoy (eds.): *WWW'07. Proceedings of the* 16th *International Conference on World Wide Web, Banff, Alberta, Canada, 8–12 May 2007*. New York, NY, USA: ACM, pp. 231–240.

Vakharia, Donna and Matthew Lease (2013). Beyond AMT: An Analysis of Crowd Work Platforms. *Computing Research Repository*, vol. abs/1310.1672, pp. 1–17.

Von Ahn, Luis and Laura Dabbish (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, pp. 319–326.

Von Ahn, Luis and Laura Dabbish (2008). Designing games with a purpose. *Communications of the ACM*, vol. 51, no. 8, pp. 58–67.

Von Ahn, Luis; Mihir Kedia; and Manuel Blum (2006). Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, pp. 75–78.

Von Ahn, Luis; Ruoran Liu; and Manuel Blum (2006). Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, pp. 55–64.

Von Ahn, Luis; Benjamin Maurer; Colin McMillen; David Abraham; and Manuel Blum (2008). recaptcha: Human-based character recognition via web security measures. *Science*, vol. 321, no. 5895, pp. 1465–1468.

Vukovic, Maja (2009). Crowdsourcing for enterprises. In *Services-I, 2009 World Conference on*. IEEE, pp. 686–692.

Wais, Paul; Shivaram Lingamneni; Duncan Cook; Jason Fennell; Benjamin Goldenberg; Daniel Lubarov; David Marin; and Hari Simons (2010). Towards building a high-quality workforce with mechanical turk. *Proceedings of computational social science and the wisdom of crowds (NIPS)*, pp. 1–5.

Wang, Jing; Panagiotis G Ipeirotis; and Foster Provost (2013). Quality-based pricing for crowdsourced workers.

Ward, Joe H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244.

Wolfers, Justin and Eric Zitzewitz (2004). Prediction markets. *The Journal of Economic Perspectives*, vol. 18, no. 2, pp. 107–126.

Yu-Wei, Chiu David Chiu (2015). *Machine Learning with R Cookbook*. Birmingham, UK: Packt Publishing Ltd.

Yuen, Man-Ching; Ling-Jyh Chen; and Irwin King (2009). A survey of human computation systems. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, Volume 4. IEEE, pp. 723–728.

Yuen, Man-Ching; Irwin King; and Kwong-Sak Leung (2011). A survey of crowdsourcing systems. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, pp. 766–773.

Zar, Jerrold H. (2007). *Biostatistical Analysis (5th Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Zhu, Dongqing and Ben Carterette (2010). An Analysis of Assessor Behavior in Crowdsourced Preference Judgments. In Vitor R. Carvalho, Matthew Lease; and Emine Yilmaz (eds.): *CSE'10. Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation, Geneva, Switzerland, 23 July 2010*. New York, NY, USA: ACM, pp. 17–20.

Apêndice A

Tarefas utilizadas nos experimentos

Os experimentos realizados para verificar o efeito do uso de testes de qualificação do tipo customizado em tarefas submetidas na plataforma Amazon Mechanical Turk fez uso de tarefas existentes na própria plataforma. Foram selecionadas tarefas das classes *Information Find* (IF), *Interpretation and Analysis* (IA) e *Content Creation* (CC), pelo fato dessas atenderem aos requisitos estabelecidos no Capítulo 5, Seção 5.1.

Para cada uma das classes selecionadas, foram escolhidas três tarefas distintas porém similares em relação aos objetivos. Além disso, testes de qualificação foram utilizados para pré-selecionar trabalhadores para a execução das mesmas, além dos outros dois cenários: qualquer trabalhador cadastrado na plataforma poderia realizar tais tarefas e, apenas trabalhadores mestres poderiam executar as tarefas.

As Figuras A.1, A.2 e A.3 apresentam as tarefas submetidas para os experimentos da classe IF, IA e CC, respectivamente.

Find e-mail addresses of American Universities

Requester: Ianna Maria SOdre Ferreira **Reward:** $0.04 per HIT **HITs available:** 0 **Duration:** 30 Minutes

Qualifications Required: Qualification test Find 1 greater than or equal to 75

HIT Preview

Instructions

Find the email address or contact form of each American University

- Go to the URL. Go to the Computer Science OR Engineering Department of the University.
- Please select which department you are in.
- Please search on the department site and find the best email address or contact form URL and paste it.

- Go to the URL. Go to the Computer Science OR Engineering Department of the University.

Website: ${website}

- Please select which department you are in. If the university doesn't have either department, choose "general".

General
Computer Science
Engineering

- Find the best email address or contact form URL and paste it.

Website address:

http://

(a) Experimento 1

Find Identify Companies/organizations from the URL's provided

Requester: Ianna Maria SOdre Ferreira **Reward:** $0.05 per HIT **HITs available:** 0 **Duration:** 1 Hours

Qualifications Required: None

HIT Preview

Instructions

Identify the company/organization associated with the provided link.

- Enter the website address for the **official website** of the company/organization.
- Include the full address, e.g. http://www.thecheesecakefactory.com.
- **Do not** include URLs to directories, general listings or anything other than the official website of the company.

Provided URL: ${website}

Website address:

http://

Submit

(b) Experimento 2

Find Names and Email Address of Article Authors

Requester: Ianna Maria SOdre Ferreira **Reward:** $0.05 per HIT **HITs available:** 0 **Duration:** 59 Minutes

Qualifications Required: Qualification test Find 3 greater than or equal to 75

HIT Preview

1. Look at the header and footer of the article for a links to email the author or to the author's profile.
2. Search Google for the author's personal website
3. Try the following searches on Google:

- "[first name]@[website they write for].com" ex: "john@life.com"
- "[first initial][last name]@[domain].com" ex: "jsmith@life.com"
- "[first name].[last name]@[domain].com" ex: "john.smith@life.com"

4. Try this tool: http://www.linksy.me.ipaddress.com(but please make sure the match it suggests make sense)

Find Contact Information for this Article's Author

Article: ${Article}

First Name*:

Last Name*:

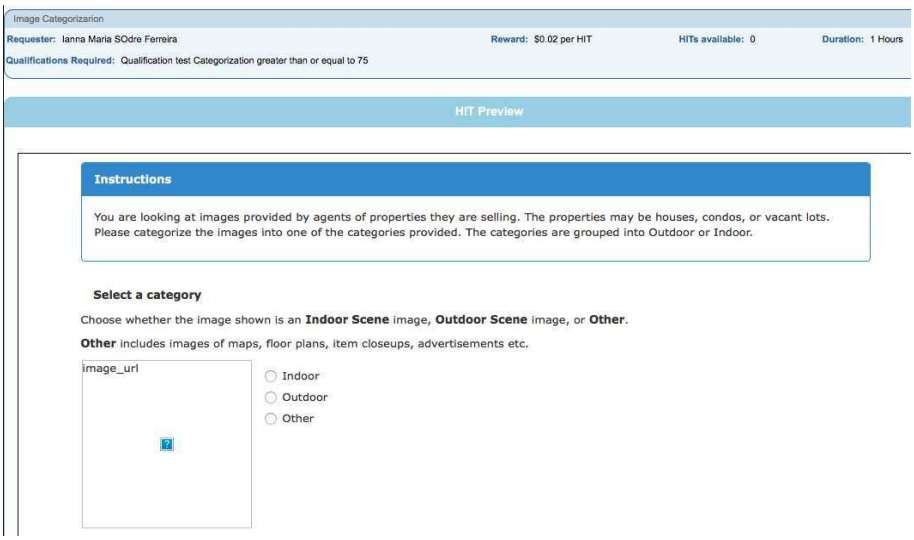
Email Address*:

Webpage where you found the contact info*: http://

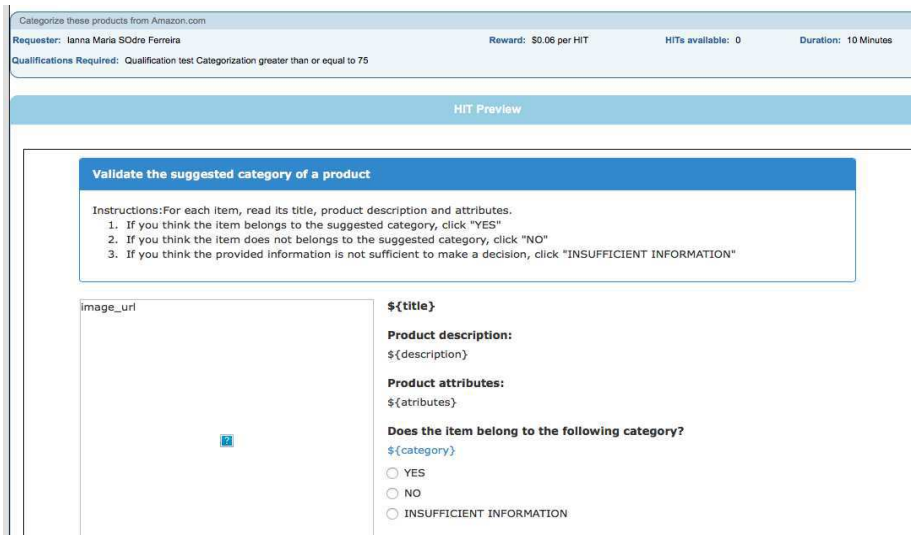
* All fields are required. Twitter/facebook pages, contact forms, etc, cannot be substituted for email address.

(c) Experimento 3

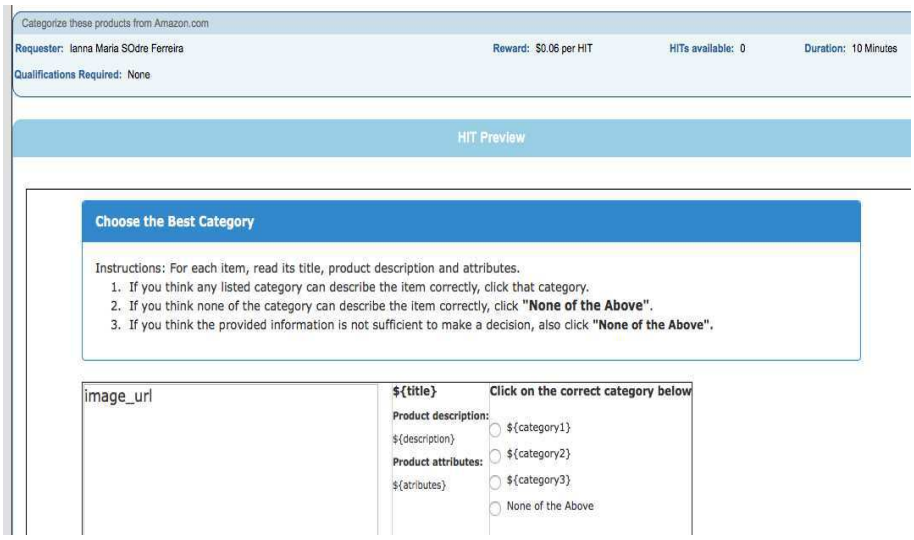
Figura A.1: Tarefas da classe *Information Find* (IF) submetidas no MTurk.



(a) Experimento 1

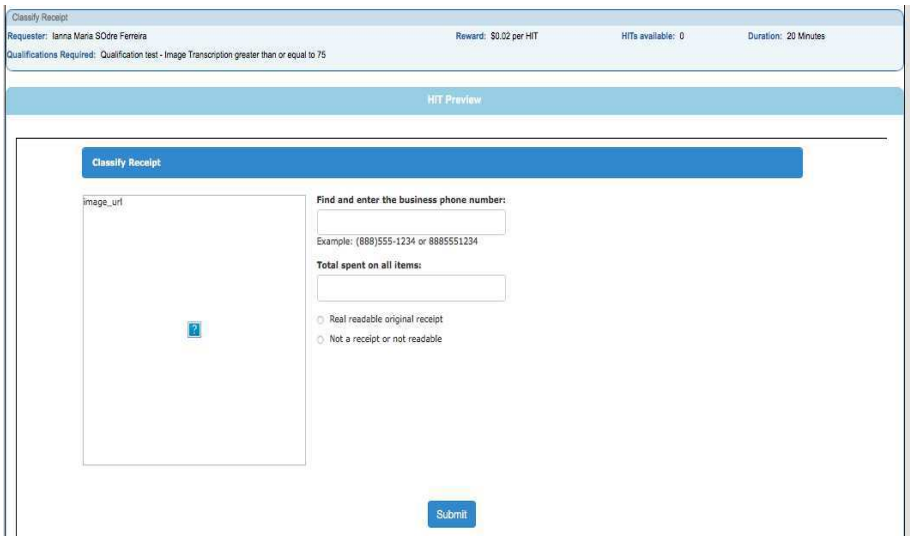


(b) Experimento 2

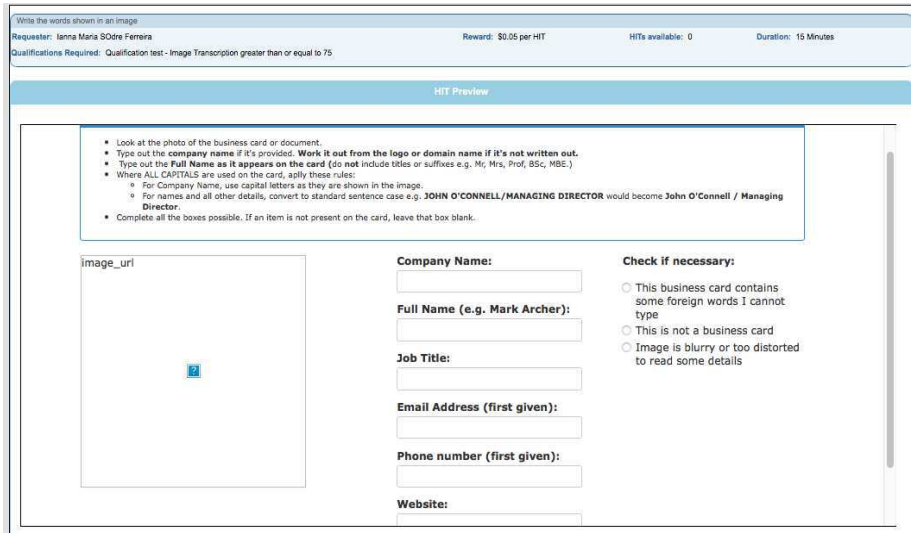


(c) Experimento 3

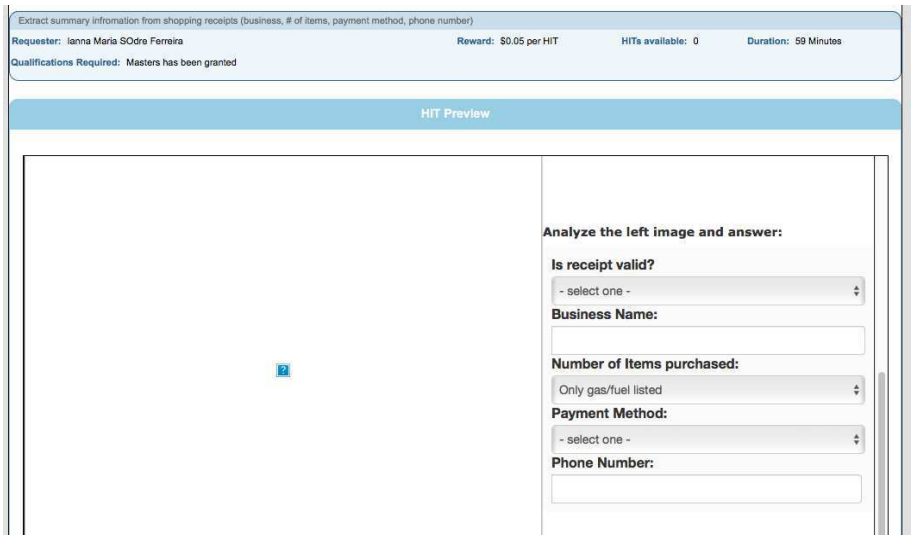
Figura A.2: Tarefas da classe *Interpretation and Analysis* (IA) submetidas no MTurk.



(a) Experimento 1



(b) Experimento 2



(c) Experimento 3

Figura A.3: Tarefas da classe *Content Crestion* (CC) submetidas no MTurk.