2021-06-15

A Framework for Assessing and Designing Human Annotation Practices in Human-AI Teaming

Suzanne Ashley Stevens
Brigham Young University

A Framework for Assessing and Designing Human Annotation Practices in Human-AI Teaming

Suzanne Ashley Stevens

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Amanda L. Hughes, Chair
Derek L. Hansen
Xinru Page

School of Technology

Brigham Young University

ABSTRACT

A Framework for Assessing and Designing
Human Annotation Practices
in Human-AI Teaming

Suzanne Ashley Stevens
School of Technology, BYU
Master of Science

This thesis work examines how people accomplish annotation tasks (i.e., labelling data based on content) while working with an artificial intelligence (AI) system. When people and AI systems work together to accomplish a task, this is referred to as human-AI teaming. This study reports on the results of an interview and observation study of 15 volunteers from the Washington DC area as the volunteers annotated Twitter messages (tweets) about the COVID-19 pandemic. During the interviews, researchers observed the volunteers as they annotated tweets, noting any needs, frustrations, or confusion that the volunteers expressed about the task itself or when working with the AI.

This research provides the following contributions: 1) an examination of annotation work in a human-AI teaming context; 2) the HATA (human-AI teaming annotation) framework with five key factors that affect the way people annotate while working with AI systems—background, task interpretation, training, fatigue, and the annotation system; 3) a set of questions that will help guide users of the HATA framework as they create or assess their own human-AI annotation teams; 4) design recommendations that will give future researchers, designers, and developers guidance for how to create a better environment for annotators to work with AI; and 5) HATA framework implications when it is put into practice.

Keywords: HATA framework, framework, human-AI teaming, artificial intelligence, collaboration, annotation

ACKNOWLEDGEMENTS

I would like to thank my husband who has been with me from the start of this degree, supporting and me mentally and emotionally especially during the writing of the master's thesis. Additionally, I'd also like to thank those who helped me during my master's thesis:

Amanda Hughes, Ph.D., for finding a project with lots of potential, helping me develop the study; and for being my friend and mentor throughout these years in the IT program;

Derek Hansen, Ph.D., for introducing me to HCI all those years go, being my professor and friend, and being a part of my committee;

Covid-19 Human-AI Teaming Project, for funding the project and providing me with the opportunity to use the research towards my thesis;

Keri Stephens, Ph.D., for her bright personality as she guided us through the interviews for the Covid-19 HAT project;

Hermant Purohit, Ph.D., for building Citizen Helper for the COVID-19 HAT project, letting us test the system, and improving it along the way;

Xinru Page, Ph.D., for being willing to jump in on my committee and offer her help and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

During a crisis event, emergency responders look for accurate and actionable data to better perform their jobs and help their communities. Social media provides a promising source of information for this purpose because people use social media during crisis events to find information, update loved ones, or call for help (Palen and Hughes, 2018; Peterson et al., 2019). For example, during the California Campfire wildfire in 2018, there were over 300k tweets about people who were potentially missing and found (Waqas and Imran, 2019).

While social media can be useful for emergency responders, often useful information is buried under massive amounts of irrelevant data (Hiltz et al., 2020; Hughes and Palen, 2012; Hughes and Palen, 2018). A popular approach to the problem of finding information in big social media data sets combines machine learning with human experts to work together as a team—we call this phenomenon human-AI teaming though it is also closely related to human-in-the-loop machine learning (Amershi et al., 2014; Stumpf et al., 2007). In this approach, programmers develop artificial intelligence (AI) systems that use machine learning to process the large streams of data, while humans provide input to the AI's machine learning algorithms to help define actionable information in the ever-changing conditions.

Human input is necessary when teaching an AI how to find relevant data, though few researchers have studied the humans that provide this input, the tasks they perform, or how they interact with the AI system to create actionable and accurate information. Stumpf believed that if

people could work hand-in-hand with the AI, then "the accuracy of [the AI] could be improved and the people's understanding and trust of the system could improve as well" (Stumpf et al., 2007). By learning more about how to better support the human annotators in human-AI teaming, we hope to improve the accuracy and efficiency of the AI.

This research studies the annotation process in a human-AI team setting and creates a framework that summarizes the factors that affect how people do this annotation work. Such a framework can help those managing the annotation process know what information and materials to include in trainings for annotators, how to reduce annotator fatigue, and what UI designs will help their annotators perform their task well. A framework can also shed light on what to look for when recruiting annotators. For researchers, this framework will provide a common vocabulary to classify, describe, and evaluate annotation work in domains that extend beyond disaster, and create a base for further studies. The framework will also be useful to developers creating software systems to support the human annotation process.

1.1 Research Questions

To better understand the human annotation process in human-AI teaming, we seek to answer the following research questions:

- **RQ 1:** What factors affect how humans annotate data in a human-AI teaming system?
- **RQ 2:** How do these factors affect human annotation?
- **RQ 3:** What are recommendations and/or design implications to improve the human annotation experience based on these factors?

1.2 Contributions

This thesis research provides the following contributions: 1) an examination of annotation work in a human-AI teaming context; 2) the human-AI teaming annotation (HATA) framework that categorizes and defines the factors that affect annotation in a human-AI team setting; 3) a set of questions that will help guide users of the HATA framework as they create or assess their own human-AI annotation teams; 4) design recommendations that will give future researchers, designers, and developers guidance for how to create a better environment for annotators to work with AI; and 5) HATA framework implications when it is put into practice.

1.3 Thesis Overview

The rest of this thesis is structured as follows: Chapter 2 discusses the research literature on social media monitoring, human-AI teaming, and frameworks. Chapter 3 captures the methodology used for studying human annotation and creating the framework for human annotation. Chapter 4 describes the HATA framework. Chapter 5 showcases framework questions, design recommendations, and framework implications. Chapter 6 concludes with a summary of the research, discusses broader impacts of the framework, and indicates directions for future research.

2 LITERATURE REVIEW

This literature review explores the history of social media crisis tools, how human experts interact with them, and existing crisis-tool frameworks that involve human-AI teaming.

2.1 Social Media Monitoring: Early Adopters

When social media first started to appear, emergency managers—who were early adopters of the technology—began to explore how it could be used in their practice. In 2011, an early case study was done on the Public Information Officer (PIO) from the Los Angeles Fire Department (LAFD) who single handedly created and incorporated social media monitoring into the department (Latonero and Shklovski, 2011). The PIO started the LAFD's Twitter account, email subscription, and text pager so emergency information would go straight to the public. Without any technical training, he cobbled together his own workflow to build keyword lists which would then be used to monitor and evaluate online content (like Twitter messages), using available off-the-shelf technologies such as Yahoo Pipes and Feed Rinse (Latonero and Shklovski, 2011). Those workflows would later become a basis upon which emergency groups could build as social media monitoring tools were created for emergency groups. Even at the beginning of social media monitoring, sifting through self-reported data was important.

As more emergency managers began to recognize the value that social media could offer response efforts, tools were specifically created to help them filter and monitor social media data.

Many of these social media monitoring tools—such as Twitcident (Abel et al., 2012; Terpstra et al., 2012), CrowdMonitor (Ludwig et al., 2015), and Ushahidi (Morrow et al., 2011)—were created by researchers at academic institutions. These tools collected social media data and attempted to filter and distill actionable insights from the data using different visualizations and presentations of the data (Cobb et al., 2014). It was then the role of the emergency manager to make sense of this data. While these tools offered substantial improvement over more manual methods of monitoring social media, they still struggled to adequately handle large volumes of data and to integrate with the workflow of emergency managers (Reuter et al., 2018).

2.2 Social Media Monitoring: AI & Human-in-the-Loop

A more recent development in addressing the problem of sifting through large amounts of data is to use machine learning that is informed by input from the humans who would use the system (Amershi et al., 2014). The AIDR system (Imran et al., 2020), as an example, uses AI to find useful crisis social media data but also uses human input to refine its algorithms. Some of the tasks that humans perform in this system include 1) gathering keywords to program the AI to gather initial data, 2) labeling the data to be used in AI training, 3) validating the labels that the AI created for the data, and 4) correcting the mapping and data visualizations created by the system. For example, a study was done during Hurricane Dorian in 2019 where digital volunteers validated labels that the AIDR image processing system had acquired and labeled (Imran et al., 2020). In this case, AIDR had already been trained, it just needed to be fed more recent data and calibrated to help with this specific natural disaster.

A similar system, Citizen Helper (Karuna et al., 2017), seeks to find actionable information for emergency responders from social media data streams. This system relies on good input from human annotators to know how to label data well. We use this system in the research proposed

here. While these systems that use human-AI teaming show much promise, we know little about how the humans that interact with these systems perform their tasks and how that affects the effectiveness of these systems for emergency responders.

2.3 Existing Frameworks with Human-AI Teaming

To help increase the effectiveness of the human-AI teaming systems for emergency responders and others, three main frameworks are of note from the research literature. One framework offers a series of sequential stages to help emergency managers sift through social media "calls for help" during a disaster (Peterson et al., 2019). The framework was meant to be used in near-real-time situations and has six stages: "1) planning, 2) collecting data, 3) filtering, 4) detecting location, 5) re-filtering based on changing disaster dynamics, and 6) sorting data into degrees of relevance to assist with emergency response" (Peterson et al., 2019). During the planning stage—stage 1—emergency managers and human volunteers decide when to collect the data for stage 2 and identify what keywords the AI should look for when filtering the data in stage 3. More direct human-AI teaming happens in stage 5 and 6, when humans re-filter the data that the computer collects. During stage 5, human annotators are looking at the tweets or images that the computer collected and deciding if each datapoint is a part of the project's goals. Everything that is a part of those goals gets passed on to help the AI learn. This stage explains the basic strategy that researchers and human annotators should follow, though it doesn't offer details on how that annotation work actually happens (which is the topic of this thesis research).

Another framework, known as the Human-AI Collaboration (HACO) framework, defines AI- and human-teaming relationships in more general terms. The framework's taxonomy considers the many roles that an AI can play in such teams and is not limited to the disaster domain: personal assistant, teamwork facilitator, associate (or teammate) and collective

moderator (Dubey et al., 2020). In turn, HACO considers many of the different ways that humans can interact with the different AI roles based on team relationships: Pure Autonomy (Human-Out-of-Loop), Teleoperation, System-Initiative Sliding, Mixed-Initiative Sliding Autonomy, and Apprenticeship (Dubey et al., 2020). The framework also includes human needs in these teams—such as Trust, Goals, and Learning Paradigms (Dubey et al., 2020)—so that humans can feel confident that they and the AI can accomplish the goals they set out to do. According to Dubey and his colleagues, some of the potential applications that the HACO framework can be applied to include On-boarding Assistance, Fashion Designer collaborations, and Cybersecurity experts and AI collaborating to detect malware (Dubey et al., 2020).

The HACO framework has much potential in the crisis informatics domain, especially when considering the human aspects— Trust, Goals, Learning Paradigms—that it addresses. The HACO framework invites people that use it to think about levels of trust so that the humans can know how much the AI will help them achieve their goals (Dubey et al., 2020). By helping the human annotators know what the AI can do, it will reduce redundancy or assumptions about the machine's capabilities. Those who use the HACO framework goals help the humans know what they and the AI should accomplish in their given task. This lessens confusion about what the goals are while also providing a means to succeed or fail. Lastly the HACO framework learning paradigms capture the mental models that the humans have "of their environment that gets updated according to events and their corresponding explanations" (Dubey et al., 2020). This idea goes hand-in-hand with trust. By understanding their environment and how they work with the AI, humans can then annotate better because they have clear goals and trust the AI system.

The human aspects that HACO framework includes can help human annotators do their task better, though there are some gaps in the HACO descriptions. First, there is little direction

on how the humans should develop trust. Identifying functionality doesn't necessarily mean that trust was created because the task was accomplished. The HACO framework suggests that goals should be made for the humans and the AI, though the framework gives no suggestions on how to prioritize the goals. While humans accomplish their tasks, they will prioritize the team goals according to their own beliefs and biases. By doing so, each human will accomplish their task differently, changing the end results of the project. Lastly, the HACO framework includes "learning paradigm" as a factor though it doesn't discuss how humans can improve their paradigm. Identification is only truly useful when action can be taken because of it. The HACO framework has another gap where it doesn't address ethical concerns with the data, such as if the data is too gruesome or vulgar for the human annotators to work with. Certain information can trigger anxiety or panic within the person while they're accomplishing the task. The mental health of the annotators is just as important as getting the task done. The end goal of human-AI teaming is having more cooperation between the human and AI so that both benefit and produce better output as a team. By identifying factors that affect annotation, and how to take action to fix the problems of how humans annotate the data, we can improve it and help others be more informed when doing annotation or designing for it.

Lastly, Barbosa et al's labeling framework uses knowledge of demographics to humanize paid crowdsourced participants by addressing two factors: *motivation* and *design*. Motivation and intention of the paid participant can greatly determine what sort of biases the AI takes in as fact. For example, a "major economic crisis in Venezuela caused many people to sign up to [a] platform in order to earn money in a more stable currency" and thus limiting the AI's capabilities to that country, culture, and language (Barbosa et al., 2019). Biases can also be introduced due to researcher's business hours and time zones, because during those times the researchers can only

reach those who are also awake and working. The framework used the example of this problem becoming biased due to gender because many "crowdsourcing platforms [have] over 70% of its available workforce as male" (Barbosa et al., 2019). By paying attention to *motivation*, researchers can be aware of potential biases and course correct the release time and date of the human-AI task in order to get more diverse participants.

The second factor, *design*, addresses the transparency of task arrangements so that participants can decide which tasks to opt-in while also allowing researchers to match up participants that would better help the AI in human-AI teams. The factor is better expressed as letting the researcher "decide how 'diverse' or 'skewed' the distribution of a certain contributor demographic must be for a given labeling task" (Barbosa et al., 2019). Essentially if the researchers notice that only Venezuelans are opting into the human-AI teaming task, researchers might consider increasing the payment amount or rewriting the task description to encourage more diversity. The design factor helps both researchers and participants in reducing unexpected tasks and controlling biases.

While the labeling framework provides great insights into paid participants, the sole two factors seem to be a condensing of factors that relate to background, training, and task interpretation factors for both the researchers and participants. The background factors for the participant might look like noting the age, language, gender, country, local economy, etc. in order to pick participants that will be better suited for the task. While those same factors are for the researchers to decide if those particular perspectives fit the overall needs of the AI. The training to better help the participants would be based on those background details, creating task instruction and rules that those people would respond to the best. And lastly, the researchers should look at the types of biases the users are bringing in, such as if people from the same place

and reason have the same inferences and if those inferences are diversified or of single mind enough to help the AI. By identifying these separate factors that affect annotation, and how they affect researchers and participants specifically, we can use that demographic knowledge to improve the experience overall when doing annotation work.

3 METHODOLOGY

The goal of this research is to better understand annotation tasks in a human-AI context and to provide artifacts (i.e., a framework) and guidance (i.e., design recommendations) for researchers, practitioners, and developers in this space. To do this, we conducted an empirical study of how humans annotate Twitter messages (tweets) in a human-AI teaming context during the COVID-19 Pandemic. After analyzing the data gathered during the interviews, we created a framework based on our observations. This chapter details the methods used.

3.1 Empirical Study of Human-AI Teaming During the COVID-19 Pandemic

During the summer of 2020, we interviewed and observed digital volunteer annotators while they annotated tweets about COVID-19 using Citizen Helper. Citizen Helper is system that finds actionable information for emergency responders from social media data streams (as mentioned in 2.2). The annotators were members of the Montgomery County CERT (Citizen Emergency Response Team). We chose annotators from this county specifically because we were annotating tweets from the National Capital Region (the Washington DC area) and their local knowledge was helpful when annotating. We refer to these annotators as digital volunteers throughout the thesis. Our community partner from Montgomery CERT, assisted in training participants to perform the annotation tasks and helped with recruiting, organizing, and

managing the annotators. Participants were paid $25 per hour for their participation in the form of an Amazon gift card.

We interviewed 15 digital volunteer annotators 3 times each. Digital volunteers annotated Twitter messages (or tweets) related to COVID-19 using the Citizen Helper tool as seen in Figure 3-1. Annotators were given a set of tweets and asked to assign the following labels (as appropriate) to each tweet: Irrelevant, Prevention, Risk, Positive Sentiment, and Negative Sentiment. A general rule was given to annotators that if a tweet was marked as "Irrelevant," no other labels should be assigned to the tweet. Though, as we'll see later in the analysis, this rule was not always followed.

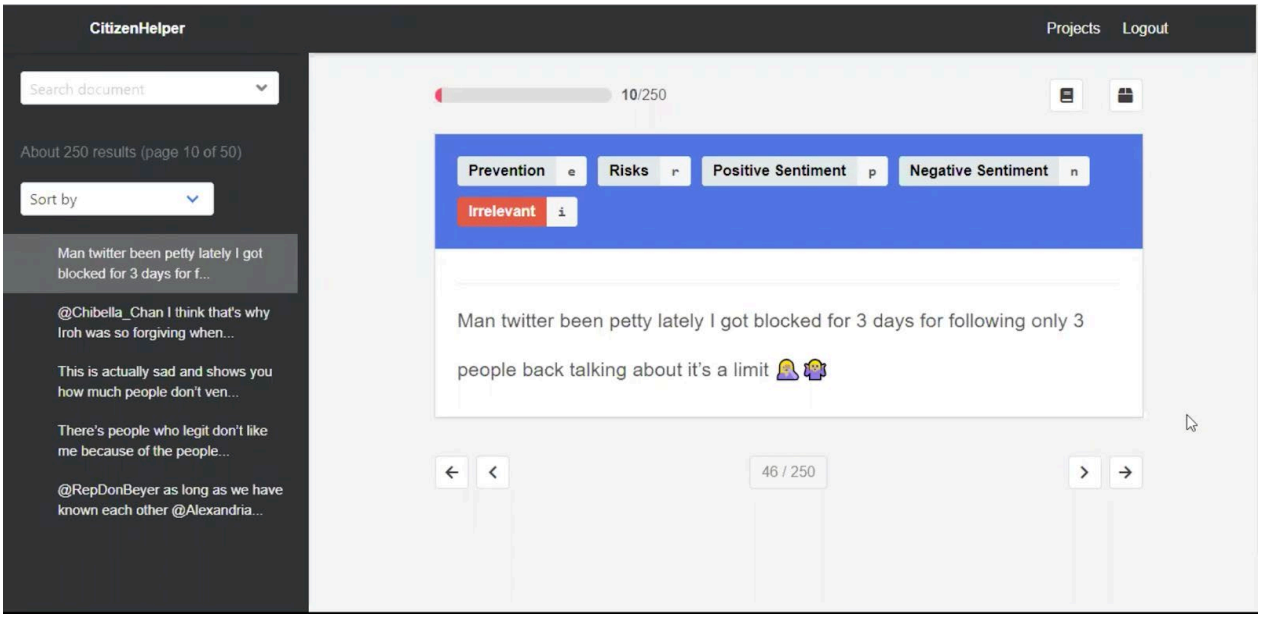

Figure 3-1: Annotating Twitter Data in Citizen Helper

All interviews were one hour in length and were conducted and recorded remotely over Zoom (a popular video conferencing service) due to COVID restrictions and the long distance between the researchers and the digital volunteer participants. In each interview, there were two

researchers present, one to lead and the other to observe. The digital volunteers were asked to share their screen so that we could see what they were seeing while they were annotating. During an interview, we periodically stopped the participants to ask them what they were thinking in the moment and why they made the labeling decisions they did. Interview strategies included cognitive interviewing and verbal protocol analysis [VPA] (Ericsson and Simon, 1984), also known as a think-aloud strategy (Lewis, 1942). We also collected digital traces of their efforts (e.g., disagreements in the human label and the algorithmically predicted labels, keystrokes, and software usage statistics) as they completed the annotation tasks.

All interviews and Zoom transcriptions were cleaned and transcribed and then loaded into Dedoose, a software for conducting qualitative analysis of textual data. The next step was to analyze and code the interview data, and then discuss emergent themes which will be discussed in the next section.

This study was conducted as part of a larger research effort led by Amanda Hughes and researchers from two other universities starting in May 2020. I led 6 interviews and observed an additional 18 (a total of 24 out of 45 interviews). I also cleaned 18 Zoom interview transcripts.

3.2 Analyzing Interview Data

Using the coding software Dedoose, we analyzed the interview transcripts to determine what factors affect human annotation. Developing the coding scheme was an iterative process, using thematic analysis techniques (Braun and Clark, 2020). To guide the analysis, we started with a few known factors that affect annotator performance such as how annotators are trained (training), how they interpret the training and data (task interpretation), and what difficulties they encounter in the user interface when labelling (UI). We built on these initial themes and identified several new ones as we read through the data and clustered similar ideas into themes.

Table 3-1 shows the major iterations of our coding scheme over time. For each iteration we would meet together and discuss and refine the coding scheme.

Table 3-1: Coding Scheme Iterations in Creating the HATA Framework

First Iteration	Second Iteration	Third Iteration	Final Iteration
Training	Training	Training	Background
Task Interpretation	Task Interpretation	Task Interpretation	Task Interpretation
UI	UI	UI	Training
	Background	Background	Fatigue
	Local Contextual Knowledge	Local Contextual Knowledge	Annotation Design
	Fatigue	Fatigue	
		Hypothetical	
		Project	
		Misc.	

Once our coding scheme was finalized, we abstracted and condensed the scheme into a list of factors and subfactors that affect human annotation. With that final list, we create a Framework (presented in the next chapter) that describes each of the factors and subfactors. Our results also include recommendations, design implications, and opportunities to improve the quality of human annotation based on the factors in the framework. This framework should have general interest to anyone engaged in research that includes human annotation of data.

3.3 Real-Time System Design

During the study, we noticed that the annotators were having problems with the task while using Citizen Helper (the system we used that finds actionable information for emergency

responders from social media data streams, as mentioned in 2.2). To help the annotators accomplish their tasks better during the study, my team met weekly on Zoom to discuss and improve the annotator experience. We discussed problems annotators encountered with the Citizen Helper interface, confusion about their task, and possible solutions. Some solutions involved changing the Citizen Helper interface, improving the training materials, and altering the instruction given to the annotators based on preliminary findings. Figure 3-2 shows the final version of the Citizen Helper interface.

We wanted this project to be responsive to the current needs of decision-makers during the COVID-19 pandemic, and not just a study that provides analysis and insight long after the event has occurred. We worked closely with our collaborator in order to do this.

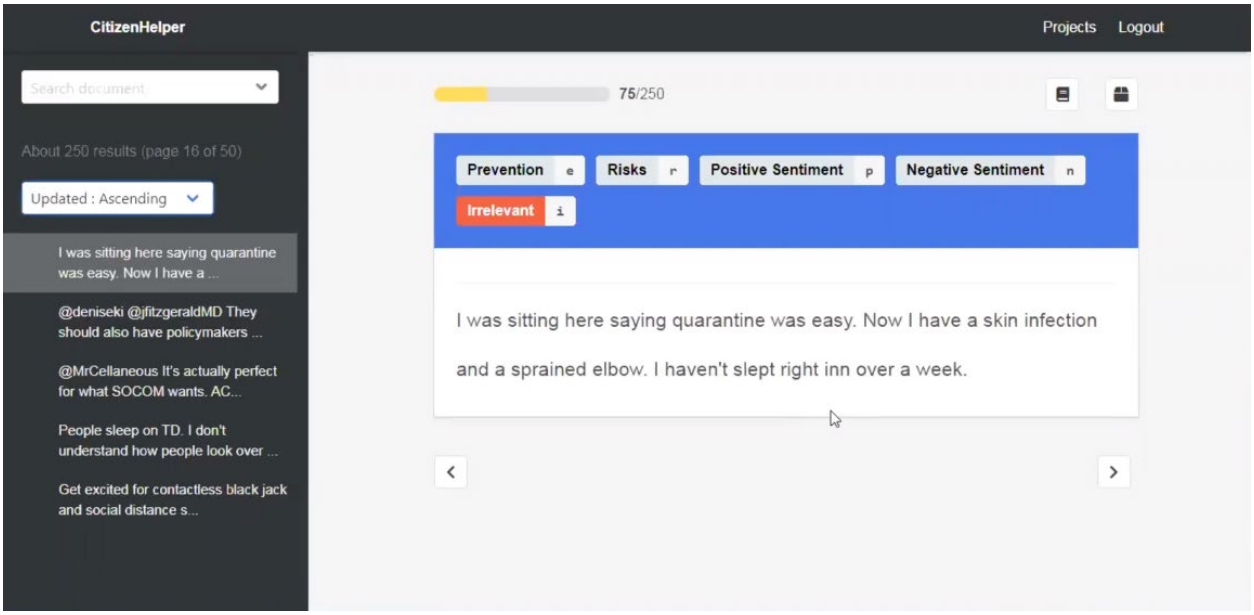

Figure 3-2: Final Version of Citizen Helper

This real-time system design was also completed the summer of 2020 as part of the larger research team. I participated in every research meeting and took part in the discussion about what should change in the Citizen Helper interface.

4 THE HUMAN-AI TEAMING ANNOTATION FRAMEWORK

In this section, we present the Human-AI Teaming Annotation (HATA) framework, which consists of five key factors that affect annotation in a human-AI context: 1) Background, 2) Task Interpretation, 3) Training, 4) Fatigue, and 5) the Annotation System. Table 4-1 defines these factors and their corresponding subcategories. The rest of this chapter will describe the framework in more detail.

4.1 Background

The background factor describes the characteristics or experiences of the human annotator that can affect how they perform their task.

4.1.1 Previous Experience with Task

People who have previously participated in a human-AI teaming project may be familiar with the type of task asked of them, or at least familiar with working with an AI. When the annotators are familiar with the teaming and task, they learn faster and have a better understanding of how the task should be done.

In our study, 7 of 15 annotators had done prior annotation work with Citizen Helper. Many

Table 4-1: Human-AI Teaming Annotation (HATA) Framework

Factors	Definition	Subcategories
Background	Characteristics of the human annotator that affect how they perform tasks	**Previous Experience with Task**—Past experience with a similar task that affects the annotator's ability to perform the current task
		Technical Proficiency—The technical ability of the annotator that helps them use the annotation system
		Contextual Proficiency—The technical ability of the annotator that helps them understand the technical context for the task
		Bias and Belief—The set of biases and beliefs held by annotators that affect how they view and perform tasks
		Past Experience—Prior life experience that shapes how annotators perform tasks
		Topical Knowledge—Knowledge relevant to the subject area of the task that can help annotators perform tasks
Task Interpretation	How the annotator understands and performs tasks	**Purpose**—The annotator's understanding of why the human-AI team was created, who benefits from the project, and how the data will be used
		AI Relationship—The annotator's understanding of how the AI works and how their contributions aid the AI
		Rules—Rules and definitions that annotators use and develop while doing the task
		Inference—The process by which annotators make conclusions about how to perform a task based on reasoning or evidence
		Perspective—The viewpoint from which the annotator chooses to analyze the data (e.g., based on the tweet's face value, author's intent, or the viewpoint of an emergency manager)
Training	How annotators are taught to complete their task	**Task Instruction**—Activities where the annotators learn the rules and goals of the project
		Resources—Materials provided to answer questions or provide guidance about the task
Fatigue	Elements of the task that stress and tire the annotator	**Task Repetition—**Performing a similar task with few variability
		Lack of Context—Missing parts of the discourse that could provide meaningful information
		Difficult Content—Content that is difficult to see
Annotation System	Aspects of the system design that affect the way humans do their tasks	**Navigation**—The ease with which the annotator can go through the system
		Task Support—System functionality that assists the annotator in performing the task

had been with the project for "three months from the initial activation" (I7), helping with a few rounds to help teach the AI initially:

> *"Covid-19 is the first time we've actually done the labeling portion of the project, but [name of community partner] and I have been working on data mining and Twitter since about 2015 or so together" (I9).*

> *"I've done some work for this project for the last six to nine months or maybe as long as a year, but I'm not quite sure" (I15).*

This helped them be more familiar with the purpose of Citizen Helper and what was required to get it to work. The other 8 annotators had other human-AI experiences where they would help identify images, medical texts, etc. One annotator said that in a previous experience, "it took us 5 million coded reports to get the NLP training right for what we were doing. So far this is a small set [for Citizen Helper]" (I7). Annotator I7 understood that it took a lot of information to teach an AI how to classify data correctly and this helped them better understand the current task with Citizen Helper and its purpose.

4.1.2 Technical Proficiency

Having annotators who know how to use the annotation system software can help improve proficiency and reduce error. An important technical proficiency for annotators also includes knowing how to work and troubleshoot the system they are using to annotate. In our study, many of our annotators had trouble navigating when moving from one tweet to another, as seen below in Figure 4-1. Many annotators would use the left sidebar to help them find the tweets that they failed to annotate due to the faulty navigation. Using workarounds, they were able to finish the tweets assigned to them. In cases where technical support was not immediately available,

annotators with troubleshooting skills found creative ways to finish the task when system problems occurred.

4.1.3 Contextual Proficiency

Annotators who are familiar with the context of the annotated data can help improve overall proficiency and reduce error. A technical context that was important for annotators in this research to understand were Twitter conventions and how they are used, such as the hashtag (#) symbol and mention (@) symbol. The hashtag (#) convention is a way for people to attach meta data to microblogs or photo-sharing services like Twitter and Instagram. People use hashtags to cross-reference content by subject or theme—e.g. *#covid19* or *#corona*. The mention (@) convention is used to tag people's accounts in microblogs or photo-sharing services. It allows people to indicate or notify certain people of their post, and those people with the *@username* who typically sent a message. Annotators who understand these symbols could not only identify who the tweet was directed to (with the @ symbol followed by a username) but could also point to tweets "where people [were] just trying to increase their presence" (I2). In contrast, annotators who did not understand Twitter conventions would make assumptions about them or not know what to make of them:

> *"Hashtags, honestly, believe it or not, even at age 66 with a lot of IT experience and computer science experience, they still baffle me" (I7).*

Having a technical proficiency with Twitter helped knowledgeable annotators understand why people used those conventions and if they would be helpful for their task.

4.1.4 Bias and Belief

We observed in our study that people were often affected in their decision-making by previously held biases and beliefs. When an annotator felt strongly about a topic, they were more likely to assign a label based on their own biases and beliefs. For example, our annotators saw a lot of tweets regarding President Trump (the president of the United States at the time of study) and many of our annotators had strong negative opinions about how the president had responded to the pandemic. One annotator was aware of her own bias and stated that she "shouldn't make a

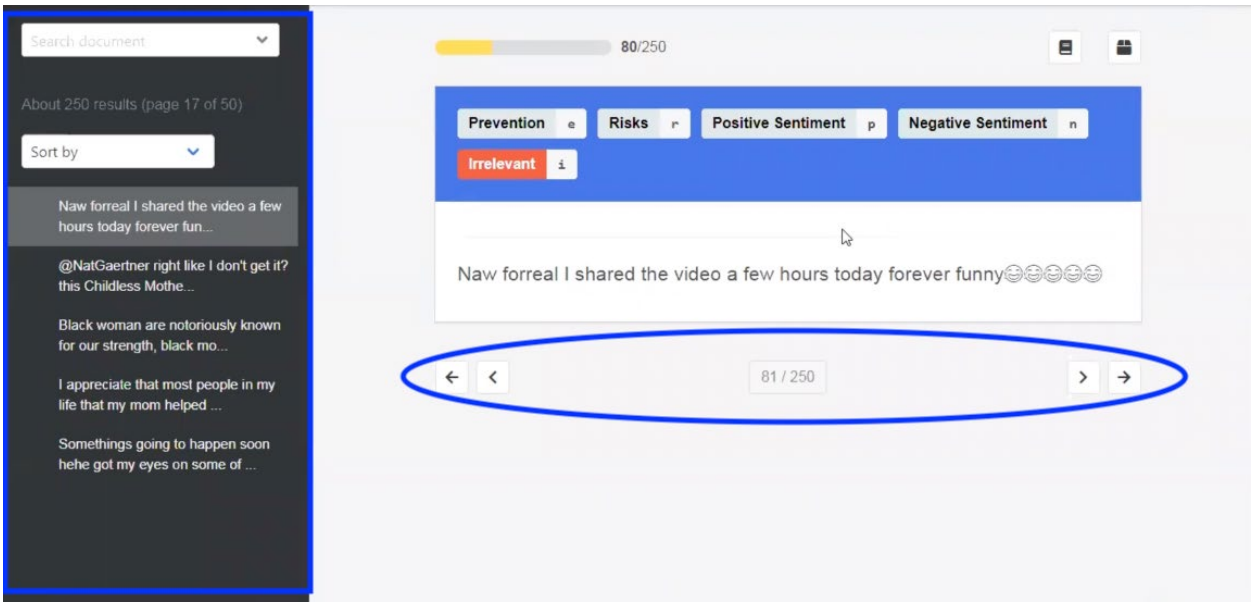

Figure 4-1: Citizen Helper Navigation Circled in Blue

judgment based on what she's seen Brennan say about Trump and the virus" (I5). Other annotators would quickly identify political tweets as irrelevant (I11).

Taking time to identify if the annotator's task involves an emotionally charged topic (like politics and religion) will help those creating the human-AI team find people who have a range of beliefs, or people who are aware of their own beliefs and strive not to have it affect their

decision-making. Also, being aware of emotionally charged topics helps the researchers not allow their own biases to creep into the interpretation of the results.

4.1.5 Past Experiences

The past experiences of annotators shape how they view and understand the task they are asked to perform. Our annotators brought with them volunteer, professional, and life experiences that affected the way they engaged with the annotation task. For example, one annotator worked in digital marketing professionally and could tell by the tweet format that it was meant to be a job posting or was tweeted by a bot:

> *"Some of the things that we see are clearly bots that are posting things, like that double line thing, that I very rarely see anybody else use besides a computer." (I9).*

Other annotators in our study were IT professionals and had some understanding of what an AI was and how to work with it. One annotator described his task with the AI this way:

> *"Well basically the way that I understand it and the way that I approach it is basically we're assisting in machine learning. We're assisting a computer to be able to make decisions about what's relevant and what's not for a specific scenario." (I6).*

He and other IT professionals thought of their task in more correct terms, which helped them be more certain of how to do the task.

4.1.6 Topical Knowledge

Knowledge relevant to the subject area of the task can help annotators perform the task better. Annotators with this knowledge can identify the right information and help the AI be more accurate. In our study it was important for our annotators to be familiar with the national capital region. Because they lived in the area, they could identify landmarks, street names, local

covid-19 restrictions, local celebrities and politicians—something that couldn't be accomplished by just anyone. During the interviews, one of our annotators stated the following:

> *"We were told there were volunteers from Brigham Young University who were also annotating. And so, a few times when I was reading [the tweets], I would see 'Bowser' and refer to DC and somebody in Utah, unless they're given a very good list, wouldn't know that's the mayor of DC" (I5).*

The annotator was right, because in another interview a different annotator informed Amanda Hughes and I that Mayor Bowser was the DC major, something we wouldn't know since we were from another part of the United States (I10).

During the study, we also found that it was important for annotators to have an understanding of general COVID-19 restrictions and health information so that they could interpret whether a tweet was relevant or not. For example, at the beginning of the study annotators notes the following:

> *"Hydroxychloroquine [was] the anti-malarial drug that President Trump was taking himself. Saying that he believed it cuts your chances of either catching COVID flat out or reduces the symptoms" (I15).*

Those who recognized the name of the drug could properly identify it as risk, since the drug could potentially cause heart rhythm problems and was shown to be an ineffective way to treat COVID-19 (Bull-Otterson, 2020). By knowing relevant COVID-19 information, our annotators could identify information that was correct or incorrect at the time of the study.

4.2 Task Interpretation

Several factors caused annotators to interpret their tasks differently in our study. These factors include: the *purpose* of the task and how it affects the goals of the project, how the

annotators view their *relationship with the AI,* the *rules* and *inferences* the annotators use while performing their tasks, and lastly, the *analytical viewpoint* of the annotators.

4.2.1 Purpose

The purpose describes the annotator's understanding of why the human-AI team was created, who benefits from the project, and how the data will be used. In our study, the annotators knew that by doing their task, they would be helping local emergency managers help those in the DC area. One annotator said that she enjoyed doing this tedious work because she knew that someday "there might be an AI program that emergency managers would use that might actually be good for something" (I11). Out of the 15 annotators, 8 others also commented on how happy they were to be a part of this project, knowing that their hard work would be put to good use helping the emergency managers help others:

> *"It is nice to see that it [machine learning and AI] has an application in emergency services and community management. I never ever would have thought to make that leap for AI to be applicable in this regard and it is very heartwarming to see that we can do some good" (I2).*

Knowing that her work with the AI would eventually be helpful made the tedious work became more worthwhile.

4.2.2 AI Relationship

When human annotators understand the capabilities of the AI and how their contributions aid the AI then the annotation task tends to go more smoothly. One annotator was so cautious that sometimes she would not "mark it [the tweet] at all. To be honest, I don't want to confuse the machine" (I2). She was so worried about confusing the AI that often she didn't include all the

tweets that were about COVID-19. She didn't understand what the AI did and therefore didn't exactly know how her actions affected the AI. She treated the AI like a child, instead of a machine and partner. Other annotators were concerned about assisting the computer and were "tuned into thinking what we wanted the computer to learn" (I11). Many of the annotators did not have good ideas of what the AI could do and so that changed how they would do their task.

4.2.3 Rules

During this study, our annotators attended a training session and where they were taught annotation rules that would the most helpful to the AI, the researchers, and eventually the emergency responders. Those rules involved labeling tweets at face value, not including sentiment if a tweet is tagged irrelevant, and the definitions of the labels to use when annotating: irrelevant, prevention, risk, positive sentiment, and negative sentiment. Tweets were labelled *preventative* if they mentioned masks, staying at home, social distancing, etc. Tweets were labelled *risk* if they mentioned no masks or ventilators, not washing hands, etc. Tweets were labelled *irrelevant* if they didn't mention preventive measures or risks, or if they weren't located in the DC area. Tweets were labelled *positive sentiment* if they were happy, joyful, optimistic, or had humor. Tweets were labelled *negative sentiment* if they were angry, hateful, pessimistic, or had worry.

Many of our annotators followed the rules well, thinking about the tweet and then applying the labels that seemed most appropriate:

> *"The irrelevant tweets that are painfully obvious in this data set. Labeling those haven't been much of an issue" (I15).*

While annotating, our annotators would sometimes forget some of the rules or apply them inconsistently. For instance, many of our annotators would apply sentiment to an irrelevant tweet after telling us that it wasn't related to COVID-19 (despite being told not to):

> *"Okay, it's irrelevant and not relevant to COVID. It's also positive sentiment at the same time, you know, helping families who have lost family members because of military service" (I12).*

The rule of not marking irrelevant tweets with sentiment might not have been as memorable or as important as others, since any tweet marked *irrelevant* wasn't viewed by the AI anyway, though placing more importance on the rule might have allowed our annotators to not spend as much time thinking about irrelevant tweets.

We found that human annotators developed their own rules to increase efficiency and make sense of the patterns they saw in the data. For example, many of our annotators decided that all tweets about politics and sales pitches would automatically be considered irrelevant. One annotator stated:

> *"I know it's talking about COVID, it's talking about the deaths, but to me it's irrelevant, because it's politically geared towards the Governor." (I10).*

Typically, annotators were more concerned about the tweet content as a whole rather than what individual parts would refer to, so after seeing a couple hundred tweets like this, many annotators formed a simplified rule where they would mark all political tweets as irrelevant. For the tweets about sale pitches, annotator 5 stated the following:

> *"Some of them were clearly sales pitches, but it seemed like there was some value in maybe 1/3 of the sale pitches. I tended to include them and now on reflection I wish I hadn't" (I5).*

This quote demonstrates how these rules about how to label certain kinds of content was an evolving process that developed over time. While completing the task our human annotators found patterns in the data and decide how to best annotate them. The larger the dataset, the more important it was for them to find ways to annotate faster and more effectively, and the informal rules they created helped them to do this.

4.2.4 Inference

Inference is the process by which annotators make conclusions about how to perform a task based on reasoning or evidence. The only information our annotators had was the training we gave them and the tweet text in Citizen Helper. We learned from an annotator that Twitter has "so many more of these conversations happening now as threading and replying", so our annotators weren't getting "the full context of [the] conversations" (I13).

To make up for the lack of context, some annotators would try to connect the dots from clues in the data. For example, one annotator inferred that Instacart would be considered relevant to COVID-19 because the people who work for Instacart:

> *"know that you have to go one way down the aisles there and they know how to be more careful and cautious. So, they're taking more preventative measures than people who aren't doing it all the time" (I2).*

She understood that if more people are ordering through Instacart for their groceries, then fewer people are going out and therefore taking preventative measures for COVID-19. That annotator made many inferences to fill in the gaps, and to finally conclude that the tweet was relevant. Not all of our annotators agreed. One annotator noted:

> *"A lot of them [other annotators] on some of the training questions [would] infer way more than I do, like way more. But I don't think that's what we're supposed to do. I infer*

some things, but I think that's the hardest part is to know when it's not black and white" (I14).

Because some of our annotators inferred more than others, they could reach different conclusion about how a tweet should be annotated.

4.2.5 Perspective

Perspective refers to the viewpoint from which the annotator analyzes the data. Some examples of different perspectives that annotators used in our study include labeling tweets based on the tweet's face value, author's intent, or from the viewpoint of an emergency manager. Annotators would choose a perspective to help them annotate the tweets more accurately. During our training, annotators were told to label the tweets at face value, as if "in a vacuum" (I7). The annotators tried to do this, but some found that it was easier for them to label the more complex tweets if they took a different perspective. One annotator created something called the "EOC test" that he would use to help him think more like an emergency manager. He would imagine:

"If I was in the EOC [Emergency Operations Center], what would I find relevant? could I take action on this information or does it help me make decisions on anything that I have going on?" (I13).

Since he was a digital marketer and not an Emergency Responder, this helped him reorient his perspective to think more like the people he was trying to help.

The different perspectives that the annotators take on are important to understand because they influence how annotators interpret the task. Annotators use those perspectives as a tool to help understand how to move forward when the rules are not clear on how to accomplish the task. In our study, we wanted to understand how our annotators were accomplishing the labeling

task and didn't anticipate some of the perspectives we would find, especially when annotator 14 talked about the persona they created:

> *"I started to create this persona because I noticed in this batch of tweets that there's a lot of tweets about a guy who doesn't believe in COVID or social distancing. Because I'm getting a lot of that kind of attitude, I created somebody who has very different beliefs and outlook than I do. I'm trying to think like him" (I14).*

This persona allowed her to better understand what the author intended with a given tweet. Though not all the perspectives that our annotators used were as helpful for the AI. As mentioned, the perspective of using the "EOC Test" that annotator 13 referred to wasn't as helpful. The purpose of labeling tweets in this round was to teach the AI to look for risks and preventative measures for COVID-19, not necessarily to find information that would be helpful for the Emergency Operations Center.

> *"So early on I found myself being more cautious and hit irrelevant, irrelevant, irrelevant on everything. And then I realized that's not doing any one a service if I'm just marking every tweet as irrelevant" (I9).*

When creating or maintaining a human-AI team, it is important to be explicit with the annotators about what perspectives are appropriate for the task. In a similar vein, team owners must be willing to find ways to alter an annotator's perspective if their labeling is putting the task further away from the overarching purpose. This way the data will be more accurate and the annotators will better understand how their perspectives impact the goals of the project.

4.3 Training

The next factor that affects human annotation in a human-AI context is training. To perform their tasks well, annotators need appropriate training instruction and access to resources

to help them understand their task. We discuss the subcategories for the training factor in more detail below.

4.3.1 Task Instruction

Perhaps the most important part of the human-AI team is defining for the human annotators what they need to do to accomplish their task. That means laying out the rules and goals of the task clearly and in order of importance, so that the human annotators know how to best spend their time. Often in our study, annotators would struggle to remember all the rules they were expected to follow. Some rules were only given as tips and tricks, so it made them seem less important. A few times during the interviews, the annotators would ask:

> *"So if I remember correctly, the protocol for this labelling was that it was prevention and risk in the DC area, right?" (I10).*

Not all of the annotators realized that if the tweet was talking about another place outside of the DC area, then it wasn't considered relevant even if it did discuss COVID-19. For example, there was a tweet about Disney Springs and 2 out of 3 of our annotators who coded it during the interviews identified it as relevant even though Disney Springs is in Florida. If the rules were defined more clearly, then our annotators might not have missed these details.

To remember the task definition while doing the task, a little practice and hands-on activities can help. In the study, our community partner taught the human annotators what to do and expect in this task. Annotator 15 who had previous experience with similar tasks recalled that they were warned that:

> *"this is a whole different ball of wax. There's all kinds of funky stuff that gets shot through and that's one of the things that [name of community partner] mentioned up front" (I15).*

During the training, the community partner would do a few questions to train the annotators with the group and ask:

> *"'what would you rate this as? and how would you do this one?' So, those sessions for me felt like they helped to prepare me better for this because then I can have a different frame of reference, as someone else that I might not have considered previously" (I2).*

For the annotators, the direction and hands-on activities helped a lot. Another annotator explained that during the training session that the community partner "only did a few in his training, he did around 15 or 20 [tweets]" (I5). The instructional strategies helped cement the task for our annotators much better than if we had just given them a sheet of instructions.

4.3.2 Resources

Resources refer to the training documents and cheat sheets that the instructors provide to help the human annotator when they are doing the task. Training only happens once or twice, but the annotator does the task multiple times, so they often need help remembering what they need to do. Many of our annotators took notes and created a list of tips of things to remember:

> *"So, in that case, I look again at my notes and those tips I told you about. I remind myself of that. I specifically asked myself, would this be useful to an emergency manager. Sometimes that helps me. I might go and look at the examples. So, in my own notes and tips from [the community partner's name] and then the examples up in the box up there [referring to the annotation guide in the UI]" (I11).*

Our community partner also gave the annotators access to the slides so that they could review when they needed to, and many of the annotators did. One annotator was "nervous about doing this [annotating the tweets] again but then I went through and reread the slideshow" (I5)

before the first interview. Having training materials and resources available for the annotators gives them confidence and guidance when approaching their task.

4.4 Fatigue

After completing the same task hundreds of times, there are elements of the task that stress and tire the annotator, making it harder for them to accomplish the task: *task repetition*, *lack of context*, and *difficult content*.

4.4.1 Task Repetition

Performing a similar task repeatedly can be exhausting. One annotator stated that "after thousands of these, I was getting a little tired of it" (I7). Across the timeframe that the three interviews took place, we asked our annotators to label 500 tweets. Even though there was a "high percentage of irrelevance" (I6), the annotators had to decide for every one of the tweets if the tweet was relevant or not. Many of the tweets were hard to decipher and would take a lot of concentration when deciding relevance. Imagine doing that same task multiple times. Eventually the stress of trying to do it right every time can wear on the annotator. While working with the annotators, our community partner mentioned that he

> *"tried to have the annotators not get so stressed out and trying to be 100% certain that they got the right labels. The reason I say that is because it just keeps the flow going and the analysis. Once you start focusing, it's kind of hard" (I1).*

4.4.2 Lack of Context

No matter the task, having a lack of context can be stressful especially when trying to make decisions. In our study, all annotators struggled with the lack of context caused by the

tweets being showcased outside of their conversation threads. Annotators could not see who posted the tweet, images or videos attached to the tweet, the date the tweet was posted, or the rest of a conversational thread that the tweet may have belonged to. This decision was made to simplify the task in Citizen Helper and to discourage spending too much time investigating each tweet. Yet, it still frustrated annotators at times:

> *"So basically, this was frustrating me. I decided I'm just marking this [tweet] irrelevant because I can't, I can't figure it out. Can't get enough information out of it to know what to do with it" (I11).*

If we had provided images or indicated that the tweet was a part of a thread (and given access to that thread), then annotators may have had enough information to label tweets that lacked context. Instead, our annotators who were aware of the other ways to get more context just had to work with what they had.

4.4.3 Difficult Content

The data that the annotators work with to accomplish their task might have difficult content. For example, during the time of our study several black people were unjustly killed by law enforcement in the U.S., which sparked much social unrest and protesting. Many people tied the killing, protests, and social unrest to the Black Lives Matter movement, which was discussed often on Twitter along with people's reactions to COVID-19. Our data included tweets about what happened around the unjust killings and COVID-19. It was difficult for our annotators to review messages about those events because they often contained accusations of injustice, racial slurs, and inflammatory language. To counteract it, our team tried to give our annotators datasets that happened before the events started. Though some of annotators still saw discussions of the events in the data, which caused confusion and stress:

"*Annotating actually was much harder after the horrible events of late in Black Lives Matter. Some of the tweets were really really hard to tell which event they're talking about and I would just make a flat judgment, basically guessing" (I5).*

Another aspect of the difficult content in the data was the profanity and racial slurs that were present in the tweets. Not all of our annotators were bothered by the profanity: some found it funny. However, the profanity did bother at least half of our annotators. One annotator mentioned that one had to "steal yourself against some of the vulgarities [because they] are pretty insane" (I11). Even though we warned annotators that they would see a lot of profanity and racial slurs, seeing them often was still stressful. To help reduce that fatigue, we replaced all profanity and racial slurs with the filter <*swear_word*> after the first round of interviews, as seen in Figure 4-2. We were also worried about the emotional wellbeing of our volunteers, especially after reading and labeling tweets for hours. The implementation picked up most of the offensive language.

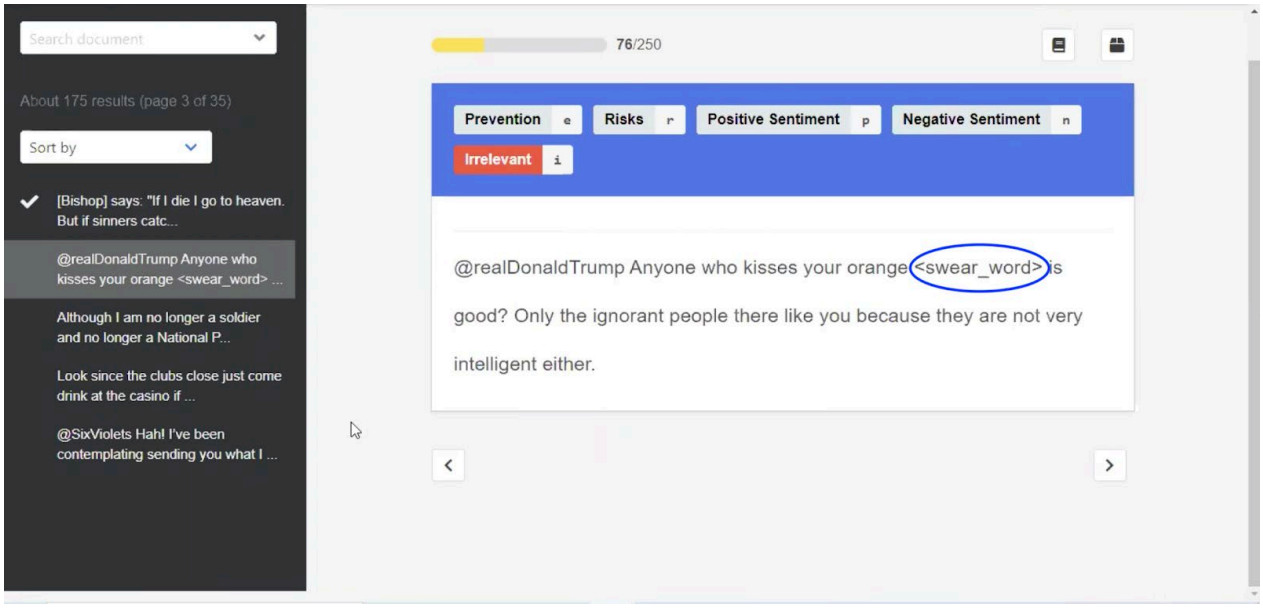

Figure 4-2: Validating Changes to the UI—remove all offensive language & racial slurs with <swear_word>.

Our volunteers had mixed reactions to the profanity filter. Half of the volunteers wanted to keep the profanity in the tweets because it helped add context. Annotator 13 noted:

> *"I think in some instances, it's hard to figure out if the <swear_word> word is a noun, verb, or adjective. And in those instances, the filter makes it more difficult to code the tweets" (I13).*

Other annotators didn't want the filter because they had "seen worse and heard worse" (I12) or because "Twitter is full of swear words" (I14). Annotator 11 decided:

> *"Taking out the swear words doesn't really help me much because whereas it might be less negative activity for me as a coder, I want to be able to be accurate. I want as much information as possible to be accurate and think about what I'm seeing. So, I think that the swear words should go back in" (I11).*

The other half of the annotators who were happy with the filter liked it because it helped them focus more on the task at hand. Annotator 6 said:

> *"the swear filter is great. It gives you less things to have to look at. It's easy to navigate the tweet without having to read a bunch of extra stuff" (I6).*

Another annotator liked that we took out the swear words because "after seeing thousands of them, I was getting a little tired of seeing the words" (I7). Other annotators were worried that the filter wouldn't catch "common abbreviations for some extremely offensive things as well" (I2) and that they would often see words "that should be added to the censoring list" (I9).

Despite the differing opinions on the swear word filter, 9 out of 15 annotators agreed that the filter shouldn't be implemented so that they could have more context to the tweets. Though if the filter must stay, then it should filter out only some words:

"There are certain swear words that are definitely more triggering than other words, like the F word or the C word or any other words that might be more triggering" (I9).

4.5 Annotation System

In this section, we discuss the aspects of the annotation system that the human annotators use to complete the task. Noted below are some of the most important aspects include *navigation and task support.*

4.5.1 Navigation

The ease with which the human annotators use the annotation system to accomplish their task is very important. Navigational issues cause annotators to not know where they are in the system or how much of the task they've completed.

In our study, our annotators had many problems understanding where they were in the system. One of the first difficulties we noticed was annotators moving from one tweet to the next. In the first interview, we had two sets of buttons and two different kinds of progress indicators (see Figure 4-3). Annotators would accidentally click the second set of buttons, essentially skipping five tweets at a time: "I like the changes where I no longer have to worry about skipping a whole five [tweets] at a time" (I11). It took a lot of mental energy for our annotators to make sure that they clicked the right button: "Whoops, sometimes that happens. I forget, and I clicked the one that takes me to the next page instead of the next tweet" (I11). After the first round of interviews, we removed the outside buttons that skipped 5 tweets forward and backward, making it easier and more enjoyable for the annotators: "I really like the way it's laid out. I really like you guys got rid of those extra buttons" (I9).

Letting our annotators know where they've been in the system greatly helped them know where they were going and how much of the task they had completed. As mentioned in Figure 4-3, Citizen Helper had two types of progress indicators. The top progress indicator showed the annotator what tweet they were on in the entire set. The bottom progress indicator represented the tweet the annotator was on in their current set. Our annotators weren't taught the difference between the indicators and were always confused as to which indicator represented what. One annotator described it best:

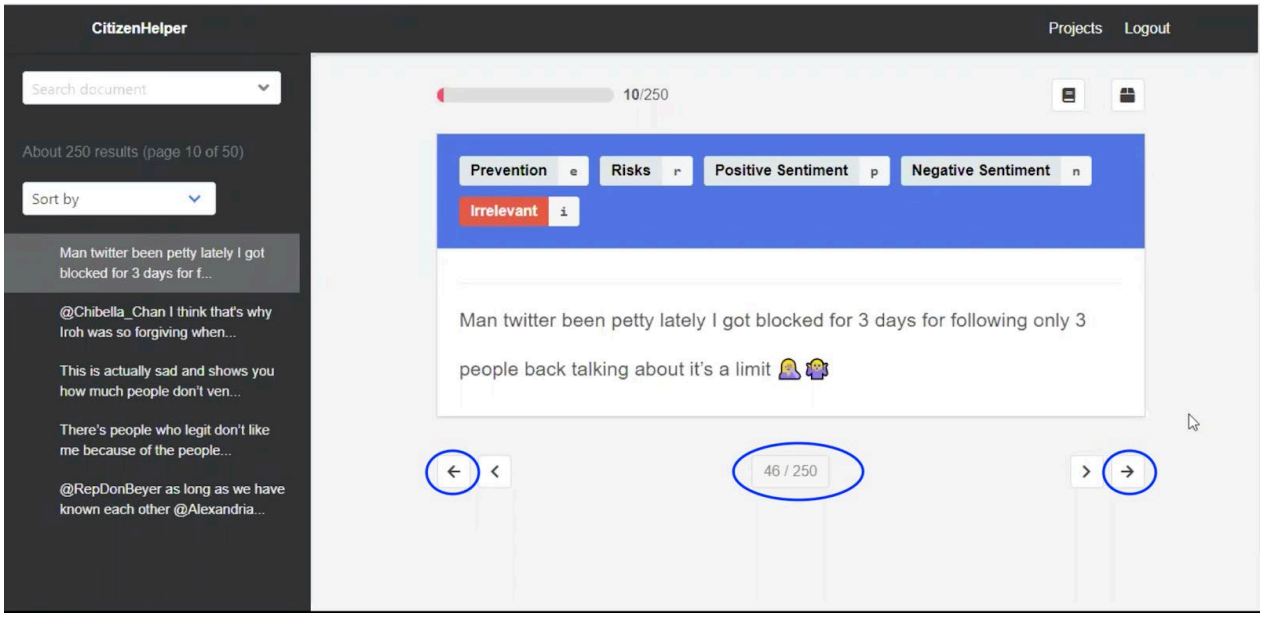

Figure 4-3: Changes to the UI—the button that skips 5 tweets forward & backward, and the bottom progress indicator.

> *"I always get lost like [up here I've] done 25 out of 250. But here [below the tweet] it says 16 over 25 [referring to the tweet he's on out of the tweets completed]. How do I get to the ones that I've completed?" (I14).*

For the second interview, we removed the progress indicator on the bottom of the screen so that no matter which tweet annotators were on, they would be able to tell how many tweets they had finished. Our annotators appreciated the change, because often they would use the indicators to tell if they had skipped tweets:

> *"I no longer have to worry about skipping a whole five [tweets]. I like that, that's a good change. And then that means I don't have to worry about those numbers. I was always comparing the numbers at the bottom with the numbers at the top because after that happened to me a few times. It's such a pain to have to go all the way back and find the one" (I11).*

One of our annotators informed us that the community partner:

> *"said that they [the system designers] had taken out the counting numbers [progress indicator] at the bottom because it was kind of a duplicate of the one at the top" (I5).*

While this explanation was not accurate (the numbers were not duplicates of one another) as to why the change took place, we learned that we needed to simplify the interface and make it more usable. As we simplified the navigation, it was easier for annotators to know where they were in the task.

4.5.2 Task Support

Task support refers to the system functionality that allows the human annotators to accomplish the task. Within our study, some of those supporting designs include displaying just the tweet text for people to annotate, easy access to labels, having a login and logout for each annotator, and providing access to the resource materials.

In Citizen Helper, the annotators navigated to a page that would show them the tweet in question with the labels in a box at the top (see Figure 4-3). Only the tweet text was shown, without any distractions. One annotator liked it because:

> *"You're not bogged down by all the other bells and whistles going off within the platform. If you're able to really focus in on just what's in front of you in the text of this particular post. You're not worried about what their full handle is, what their profile picture might be, how many tweets or retweets or likes it has. You're able to focus on the gist of it and tease out what actually is relevant to what we want to find" (I13).*

The simplification helped the annotators focus on only the text, reducing distractions from parts of the platform that would confuse people who weren't familiar with Twitter.

In our study, we tried to simplify the annotation process as much as possible. Above the tweet, the labels were neatly set in a row, always in the same order, and each label had a letter next to it indicating a keyboard shortcut. Using the shortcut or "hotkeys and your arrow keys makes it a lot easier for us and [makes it] faster to navigate through the whole process" (I15). The *irrelevant* label was red to help annotators identify it easily: "I like the fact that irrelevant is highlighted in red" (I7). By creating supporting system designs like this, the annotation process can go faster and have better results.

Another supporting system in the study was the Citizen Helper login. It helped separate different batches through a login, which we sent via email with the new username and password for every batch. We had some problems with using multiple usernames and passwords per annotator. Sometimes the annotators would lose the email or "hit delete on all of his [name of community partner's] emails" (I14). While the supporting system can be improved, the login did separate the batches properly and sent the right data to the right annotator.

Including help resources in the system design allows human annotators easy access to the materials without having to spend much time looking for the information. In our study, we did include an annotation guide within the system design though only 8 out of 15 annotators even knew it was there. The rest of the annotators sifted through their "own notes and tips from [name of the community partner] and then the examples" (I11) as a reference. Though sometimes the printed notes would be in large piles (I11) and the available slides from the training would be rather long, in our case it was "67 pages long, 67 slides" (I6) which is a lot to go through. The annotation guide on the other hand was short and often times used as a refresher. One annotator liked "having it there, rather than me having to print it out and keep it on my desk, which looks like a disaster zone of itself" (I9). Having a short guide of how to complete the task in the system design provided support, helping annotators find the right answers within the system.

4.6 Summary

This chapter outlined the HATA (Human-AI Teaming Annotation) framework, describing in detail the five key factors and subcategories that affect human annotation: 1) Background, 2) Task Interpretation, 3) Training, 4) Fatigue, and 5) the Annotation System. Background described the characteristics of the annotator that affect how they perform their task. Task Interpretation described how the annotators understand and perform their tasks. Training described how annotators are taught to complete their task. Fatigue describes elements that make the task more exhausting for the annotator. And lastly Annotation System describes aspects of the system design that affect how humans complete their task.

The next chapter will describe the HATA framework implications when it is used in a real human-AI team environment.

5 HATA FRAMEWORK IMPLICATIONS

In this chapter, we will discuss the implications of the HATA (human-AI teaming annotation) framework. We begin with a set of questions that users of the framework can use to better understand and design for a human-AI annotation team. We then offer design recommendations of the framework, followed by a discussion of how the framework can be applied and possibly extended to other domains.

5.1 Framework Questions

When seeking to create, understand, evaluate, or improve a human-AI annotation team, it can be useful to step through a set of guided questions. Here in Table 5-1, we provide a list of questions based on the HATA framework than can help researchers, team managers, and system designers know what to consider when working with human-AI teams doing annotation work. The following questions (see Table 5-1) are listed by the factors and subfactors as displayed in the HATA framework.

5.2 Design Recommendations

Based on this study, the following design recommendations have been derived from the factors and subfactors of the HATA framework. The implementation of each recommendation will need to be tested in the future to determine the usefulness of each recommendation.

To help improve the annotator's *Technical Proficiency*, there are a few changes one can make to the system design regarding the item being annotated, documentation to help the annotator use the system, and reporting functionality. Regarding the item being annotated, the system designer can add hints or documentation to help the annotators understand what each

Table 5-1: Framework Questions

Factor	Questions
Background	• What kind of characteristics do you need in your annotators? • Do they need a particular skill set? • Have your annotators worked on this project or a similar one before? • What beliefs or biases would throw off the data? • Does your group have varied enough backgrounds? • What annotator knowledge will be important for the project?
Task Interpretation	• What is the purpose of the system? • What will the AI contribute? • What will the people contribute? • How will understanding the people's relationship with the AI improve their contributions? • What patterns might annotators see in the data? • Should inference be reduced? If so how? • Should annotators pick a perspective to analyze the data with?
Training	• What kind of training will the annotators need? • Are there some activities that will improve their understanding of the task? • What rules are the most important for annotators to know? • Are there any unspoken rules that are not listed in the labels? • What materials and resources should be provided to annotators?
Fatigue	• What is stressful about the task? • What about the task might be confusing? • When will annotators get a break from annotating? • How will you encourage annotators to take frequent breaks?
Annotation System	• How complex is the navigation system? • Are there any repetitive navigation elements? • How will annotators know where they are in the system & how to move forward with their task? • Are the support functions self-explanatory? • Is the login process simplified? • Is there a place for annotators to give feedback? • Can annotators access the project rules without leaving the system?

item means. For example, in our study, our annotators struggled to understand what a hashtag (#) or a mention (@) was. To help improve this misunderstanding, when users hover over the hashtag or mention we could provide a tooltip pop up that defines the item. We could also include that documented information for the annotators, along with videos and suggestions of what to do if something goes wrong in the system. Another design improvement would be to include reporting functionality that annotators can use if something goes wrong. Oftentimes problems happen while in use and developers can't always predict all the ways a system can fail.

For *Biases and Beliefs* and *Topical Knowledge*, there are a few design changes that we could make to help improve the annotator experience. If the task involves controversial topics (religion, politics), the system could warn the annotators to prepare them. Also, there could be an information section that give suggestions on how to be aware of one's own bias. For *Topical Knowledge*, if important to the task, extra information the annotator needs could be built into the annotation system. In our study, we could have included a label for *DC Area* which would remind our annotators that that's something we wanted them to look for.

Design recommendations for the *Task Interpretation* subfactors involve providing more information for the annotators. To help remind the annotator of their *Purpose* and *AI Relationship*, the annotation system should include information about the project goal and an explanation of how the annotator will interact with and contribute to the AI. Another way to help improve the AI Relationship would be to report how the annotator or the group of annotators are helping the AI. The more involved in the process, the easier it is for the annotators to understand how their task is impacting the AI. To help the annotators understand the annotation *Rules* and control their *Inferences*, the rules could be mapped out in a decision tree for the task. This would help to systemize the process and guide annotators during decisions making with digestible steps.

It'll also reduce the number of inferences that annotators make because they would have more guidance about how the task should be completed.

To help remind annotators about the *Task Instruction* they received during their training, providing condensed *Resources* on the platform will allow them to view training and help materials inside of the annotation system. Showing up as a modal window, these resources would act as a cheat sheet, including definitions of labels and examples from the training. Training information should be provided in a more extended format outside the annotation system (e.g., a website) for annotators who wish to revisit what they learned in the training.

Reducing the *Fatigue* of annotators is important. A few design suggestions include giving annotators small batches of data to annotate at a time, thus reducing the *Task Repetition* to a reasonable amount. For example, in our study we gave our annotators 500 tweets to annotate during the 3 interviews. Instead, we could have given the annotators multiple batches of 50 or 100 tweets. Then if they finish the batch quickly, they could open the next batch to complete. The smaller grouping would make the task more manageable to complete and would give and annotators time to rest in-between if they wish. Of course, 50 or 100 tweets might be too little, and a good number of tweets would likely need to be experimentally developed. It is also likely that the optimal number of tweets to label in a batch might vary by person or circumstance.

Another suggestion to decrease exhaustion would be to ask the annotator to define how long they want to annotate before taking a break, say between 20–60 minutes. Then when the timer goes off within the system, a window would pop up and require the annotator to take a 3–5-minute break before starting the process again.

A final design recommendation in this category would be to include a way for annotators to report important or problematic content that they encounter. In our study, the swear word filter

missed some words and if we allowed the annotators to tag them, then we could use that data to improve the system. Also, during the study, annotators worried about what they should do if they found personally identifiable information (e.g., the address or phone number of a vulnerable person) or information that they thought emergency responders or local authorities should immediately see. By allowing annotators to report information like this, it would provide them opportunities to improve the system and to feel like they are helping.

The most important design recommendation for the *Navigation* subfactor in the HATA framework is that the annotator should know where they are within the system and how to navigate through the system. This means if the task involves annotating texts or images then there must be a progress indicator notifying the user which text or image they are currently on and how many remain. There should also be a simple way for the annotator to move forward and backward between the data. The simpler the navigation, the better.

5.3 Framework Implications

This thesis presents the HATA framework which defines and categorizes the factors that affect human annotation within a human-AI team. The framework can be used to design new human-AI annotation teams as well as assess existing human-AI annotation teams.

By using the HATA framework, those teams can create more accurate results all while providing a better experience for the annotators. Researchers can use the Background questions to identify potential background factors that might affect their team when looking for annotators so to best match the task to the right kind of individual. For example, if the task involves political unrest it'll be important to ask potential annotators about their political beliefs, how involved they are in their local politics, and if they've done similar work with AI before. It'll also be important to find a variety of people to not bias the results of the task.

Those who create tasks for annotators should identify the purpose of each task in the context of the project, how the AI learns and works with the annotator, and how annotator interaction with AI will make the task easier in the present or near future. Task creators should also watch for patterns in the data that annotators bring to their attention. The assumptions annotators draw about these patterns will affect the result of the project. In our political AI team example, if all the annotators claim that every tweet made by a certain political party is irrelevant, despite the task, then the results will be skewed towards one set of political beliefs. This is where good training and resources will help, especially after making sure to select a wide variety of annotators.

Task creators can also use the questions based on the HATA framework to address potential ethical concerns ahead of time. For example, they can build in mechanisms to combat annotator fatigue (e.g., limiting the size of datasets, reminders to take breaks) or ways to warn about or hide objectionable content. In turn, this could help those in the human-AI teaming field become more ethically minded when choosing tasks for annotators to do. Lastly reducing vagueness and complexity in the annotation system by using the HATA questions will help annotators be less frustrated by the process. Bias and fatigue tend to come into play if the system design is too difficult to use.

5.4 Summary

This chapter outlined the HATA (Human-AI Teaming Annotation) framework implications by giving a list of questions for creators and designers of Human-AI annotation teams to ask themselves for every factor category, a list of design recommendations that will improve the annotation system based on the framework, and lastly possible framework implications when it is put into use.

6 CONCLUSION

In this final chapter, we provide a summary of the project, discuss broader impacts of the framework, and future research opportunities.

6.1 Thesis Summary

The purpose of this thesis work is to learn more about how human annotators accomplish tasks in a human-AI teaming context. The research project consisted of an empirical interview study of 15 human annotators in the DC area. For the study, we explored how people annotate Twitter messages (tweets) in a human-AI teaming context during the COVID-19 Pandemic (Chapter 3). Through the interviews, we sought to understand how annotators accomplished their task and how people worked in human-AI teams.

Through analysis of the interview data, I developed the HATA (human-AI teaming annotation) framework (Chapter 4). This framework provides five key factors that affect annotation in a human-AI context (RQ1). I then described how each factor affects human annotation (RQ2). Finally, I provided design recommendations and implications based on the framework (Chapter 5, RQ3).

6.2 Broader Impacts

The HATA framework was designed based on the experience of human annotators working as part of a human-AI team, where they generated the training data that the AI uses to improve itself. The end goal of the broader project (outside the scope of this thesis) was to develop machine classifiers that sift through social media data to identify information that would be helpful for emergency responders. However, the framework developed here can also contribute to broader knowledge of how to help people accomplish tasks (tasks beyond annotation) in human-AI teams, as well as offer insight about applications outside of the emergency response context.

Regardless of the task, creators of any human-AI team will create a better team if they understand the background of the people they recruit and whether they will be a good fit for the task. It's unlikely that every volunteer who becomes an annotator will be in the same place at the same time. And it's even more unlikely that the trainer will be available when the annotators are working with the data to help them and solve every technical and task-wise problem that they encounter. So, it's useful to have annotators with enough technical proficiency to use the system and solve problems on their own. Topical knowledge and knowing people's past experience will help those who want to create any form of a team to find a good range of people who understand and can contribute to the topic. Understanding the background of people and matching them to the task and intended results will help not only create a better team, but a better result as well.

Understanding that people are humans and can experience fatigue when completing tasks can apply to many types of teams, human-AI related or not. As we found in this study, people tire, and some kinds of data or tasks can be distressing. Addressing ahead of time the stressful elements of a task and identifying mitigation plans can help build better relationships between

task creators and the volunteers who complete the tasks, as well as foster healthier mental spaces for volunteers.

The factors included in the HATA framework are broad enough to be used outside of the emergency response domain. There are a multitude of human-AI teams that would proffer from understanding their annotators better. For example, if a type of human-AI team were put together to encourage the public to verify if AI data was correct on documents, knowing what would interest the annotator to do the task would be incredibly important, along with providing the necessary tools if no training was allowed. The HATA framework would also be useful with tasks like identifying misinformation in social media, identifying people or objects in photos, creating a better environment in customer service or education, or learning how to have a conversation with someone through Siri or Alexa. Computers are so heavily involved in our lives that if every interaction in every domain paid attention to human needs, communication and coordination would become much simpler.

6.3 Future Research

The HATA framework is based on one empirical study of human-AI Teaming during the summer of 2020, using only one type of human-AI interaction. The next logical steps would be to validate the HATA framework through literature and testing validation.

A first step would be to do a more detailed comparison of the HATA framework with other frameworks currently in the literature. The HACO framework (Dubey et al., 2020) and "call for help" framework (Peterson et al., 2019) mentioned in the Literature Review (section 2.3) are similar to the HATA framework and share some important characteristics, yet they also have different purposes. By comparing them, we can see what new insights the HATA framework

contributes to our understanding of annotators and annotation tasks, and perhaps identify areas that we may have missed.

Another next step would be to test the framework and its various factors and subfactors to see what tradeoffs they present to annotator wellbeing, efficiency, or the quality of the task results. This could be done by creating two human-AI projects, one experimental group created with the HATA framework questions in mind for one factor and the other the control group. An example study might test the fatigue levels of two groups of people as they asked Alexa (the Amazon virtual assistant AI) to set an alarm for the next morning. We would gather volunteers from multiple backgrounds and dialects of English for this study. One group would be the control group and would ask Alexa to set an alarm. The experimental group would ask another AI which had been previously modified to understand when someone is frustrated or fatigued and change accordingly. In this example, the factors and subfactors that are most important were Background and most of its subfactors (excluding Bias & Belief), Task Interpretation (Purpose, AI Relationship, and Inference), Fatigue (Task Repetition), and Task Support. Of all the factors, the most important involved Fatigue and Task Repetition, since that was the main factor that was being studied. Reducing Fatigue in the Alexa example would create better end results for the project and volunteers. However, if the researchers focused on AI Relationship, they might have focused more on the personality of Alexa rather than her language processing or protocols to identify frustration While testing in this Alexa example, researchers might find other factors that are more important than the ones listed in the HATA framework.

Next, testing the HATA framework against different participant types and if the framework still works with participants with different motivation levels, as addressed by Barbosa et al. (2019). The HATA framework was created after observing volunteers who were handpicked by

our community partner and had been volunteering in the CERT for years or decades. Those people were dedicated to doing the task that was assigned to them and would finish no matter what. Results of the effectiveness of the framework might change if used on different types of people like crowdsourced volunteers or paid participants. There might be factors that work better with some groups than others. For example, what might be considered a decent and ethical pay to one group might have to be addressed in the Background factor for HATA along with different community motivations.

Another step would be to test the HATA framework against other types of human-AI systems and tasks. According to the HACO framework, there are many different types of roles that the AI could play in the task (such as personal assistant, teamwork facilitator, associate or teammate, or collective moderator), and different types of team relationships people and AI could have with each other (pure autonomy, teleoperation, system-initiative sliding, mixed-initiative sliding autonomy, or apprenticeship) (Dubey et al., 2020). By applying the HATA framework in those different situations, we can see how well it works and what factors would need to be improved. Another way to test the HATA framework would be to use it in different tasks. In our study, we asked our annotators to label Twitter data. Future research might ask whether the HATA framework applies well to tasks like image recognition, translating services, or providing customer service, and what adjustments to the framework (if any) need to be made to accommodate a broader range of tasks.

REFERENCES

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., & Tao, K. (2012). Semantics + filtering + search = twitcident. Exploring information in social web streams. *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, 285–294. https://doi.org/10.1145/2309996.2310043

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine*, *35*(4), 105–120. https://doi.org/10.1609/aimag.v35i4.2513

Barbosa, N. M., & Chen, M. (2019). Rehumanized Crowdsourcing: A Labeling Framework Addressing Bias and Ethics in Machine Learning. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300773

Braun, V., & Clarke, V. (2012). Thematic Analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 57–71). American Psychological Association. https://doi.org/10.1037/13620-004

Bull-Otterson, L. (2020). Hydroxychloroquine and Chloroquine Prescribing Patterns by Provider Specialty Following Initial Reports of Potential Benefit for COVID-19 Treatment—United States, January–June 2020. *MMWR. Morbidity and Mortality Weekly Report*, *69*. https://doi.org/10.15585/mmwr.mm6935a4

Cobb, C., McCarthy, T., Perkins, A., Bharadwaj, A., Comis, J., Do, B., & Starbird, K. (2014). Designing for the deluge: Understanding & supporting the distributed, collaborative work of crisis volunteers. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '14*, 888–899. https://doi.org/10.1145/2531602.2531712

Dubey, A., Abhinav, K., Jain, S., Arora, V., & Puttaveerana, A. (2020). HACO: A Framework for Developing Human-AI Teaming. *Proceedings of the 13th Innovations in Software Engineering Conference on Formerly Known as India Software Engineering Conference*, 1–9. https://doi.org/10.1145/3385032.3385044

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data* (p. 426). The MIT Press.

Hiltz, S. R., Hughes, A. L., Imran, M., Plotnick, L., Power, R., & Turoff, M. (2020). Exploring the usefulness and feasibility of software requirements for social media use in emergency management. *International Journal of Disaster Risk Reduction*, *42*, 101367. https://doi.org/10.1016/j.ijdrr.2019.101367

Hughes, A. L. (2020). "Lessons from Observing Human Tweet Annotation during COVID-19." *Text Retrieval Conference (TREC) 2020*, *TREC-IS Track.* Virtual conference, Invited keynote talk.

Hughes, A. L., & Palen, L. (2012). The Evolving Role of the Public Information Officer: An Examination of Social Media in Emergency Management. *Journal of Homeland Security and Emergency Management, 9*(1). https://doi.org/10.1515/1547-7355.1976

Imran, M., Alam, F., Qazi, U., & Peterson, S. (2020). Rapid Damage Assessment Using Social Media Images by Combining Human and Machine Intelligence. 13.

Karuna, P., Rana, M., & Purohit, H. (2017). CitizenHelper: A Streaming Analytics System to Mine Citizen and Web Data for Humanitarian Organizations. *ICWSM*. https://pdfs.semanticscholar.org/4186/97327004b486941c85e18a164b246b746459.pdf?_ga=2.183936160.784070861.1602193783-1491622627.1602193783

Latonero, M., & Shklovski, I. (2011). Emergency Management, Twitter, and Social Media Evangelism: International Journal of Information Systems for Crisis Response and Management, *3*(4), 1–16. https://doi.org/10.4018/jiscrm.2011100101

Lewis, C. (1982). Using the Thinking-Aloud Method in Cognitive Interface Design (IBM Research Report RC 9265 No. 9265).

Ludwig, T., Reuter, C., Siebigteroth, T., & Pipek, V. (2015). CrowdMonitor: Mobile Crowd Sensing for Assessing Physical and Digital Activities of Citizens during Emergencies. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 4083–4092. https://doi.org/10.1145/2702123.2702265

Morrow, N., Mock, N., Papendieck, A., & Kocmich, N. (2011). Independent Evaluation of the Ushahidi Haiti Project. Development Information Systems International. https://www.researchgate.net/publication/265059793_Ushahidi_Haiti_Project_Evaluation_Independent_Evaluation_of_the_Ushahidi_Haiti_Project

Hughes, A. L., & Palen, L. (2018). Social Media in Disaster Communication. In H. Rodríguez, W. Donner, & J. E. Trainor (Eds.), *Handbook of Disaster Research* (pp. 497–518). Springer International Publishing. https://doi.org/10.1007/978-3-319-63254-4_24

Peterson, S., Stephens, K. K., Hughes, A. L., & Purohit, H. (2019). When Official Systems Overload: A Framework for Finding Social Media Calls for Help during Evacuations. *Proceedings of the 2019 Information Systems for Crisis Response and Management Conference* (ISCRAM 2019). http://idl.iscram.org/files/stevepeterson/2019/1928_StevePeterson_etal2019.pdf

Reuter, C., Hughes, A. L., & Kaufhold, M.-A. (2018). *S*ocial Media in Crisis Management: An Evaluation and Analysis of Crisis Informatics Research. *International Journal of Human–Computer Interaction, 34*(4), 280–294. https://doi.org/10.1080/10447318.2018.1427832

Stumpf, S., Rajaram, V., Li, L., Burnett, M., Dietterich, T., Sullivan, E., Drummond, R., & Herlocker, J. (2007). Toward harnessing user feedback for machine learning. *Proceedings of the 12th International Conference on Intelligent User Interfaces*, 82–91. https://doi.org/10.1145/1216295.1216316

Terpstra, T., de Vries, A., Stronkman, R., & Paradies, G. I. (2012). Towards a RealtimeTwitter Analysis during Crisis for Operational Crisis Management. *Proceedings of the Information Systems for Crisis Response and Management Conference* (ISCRAM 2012).https://www.researchgate.net/profile/Teun_Terpstra/publication/260123055_Towards_a_realtime_Twitter_analysis_during_crises_for_operational_crisis_management/links/00b4952fa3bfc58ca2000000/Towards-a-realtime-Twitter-analysis-during-crises-for-operational-crisis-management.pdf

Waqas, H., & Imran, M. (2019). CampFireMissing: An Analysis of Tweets About Missing and Found People from California Wildfires. *Proceedings of the 16th International Conference on Information Systems for Crisis Response And Management,* 9. https://mimran.me/papers/CampFireMissing_An_Analysis_of_Tweets_About_Missing_and_Found_People_ISCRAM2019.pdf

APPENDIX A. PROTOCOL FOR INTERVIEW #1

A. INTRODUCTIONS AND FIRST CODING SESSION

A.1 Overview

In the first of three interviews, the interviewers will introduce the interview process, meet interviewees, and familiarize themselves with the coding task. The interview will take place via Zoom, with participants sharing their screens, so interviewers can see their screens as they complete their tweet labeling task. The interview should take no longer than 1 hour.

A.2 Informed Consent

[Script] Thank you so much for allowing me to sit in on your coding session today and ask you questions about your process. I'd like to verify that you have read the informed consent form and are OK participating in our study. I will also assign you a participant number, so we can protect your privacy and confidentiality in the study.

Make sure to change Zoom name to participant number and press record.

A.3 Opening Questions

Only ask these questions at the start of the first interview with the participant.

1. How would you describe the type of coding work that you do as a volunteer?
2. How long have you been doing this online coding work?

3. Have you done online coding work with disasters other than COVID-19? If so, which ones.

A.4 Tweet Labeling Task Questions

Ask these questions, when appropriate, during each interview session.

At this time, we will ask them to start their coding and speak out loud as they think through their coding decisions. These are the probing questions we will use to better understand their decision making.

1. Describe why you decided to code that tweet in this particular way?
2. Have you seen tweets like this before?
3. Is this tweet unique to COVID-19?
4. Was that a difficult tweet to code, and if so, why?
5. Was that an easy tweet to code, and if so, why?

A.5 Post-Session Questions

At the end of the session (at 45 minutes), ask these questions:

1. Now that you have finished this session, how well do you think you were able to code these tweets?
2. How well do you think other CERT volunteers might do if they had been coding this same content?
3. What types of tweets do you think a computer could automatically code?
4. What types of tweets do you think that the computer would not be able to code?
5. What do you believe is the biggest value that you bring to helping the computer better learn how to automatically label tweets?

6. Is there anything I haven't asked you that you think might help me better understand how you worked today to accomplish your tasks?

APPENDIX B. PROTOCOL FOR INTERVIEW #2

B. CITIZEN HELPER INTERFACE & PERCEIVED COMPUTER REASONING

B.1 Overview

In the second of three interviews, interviewers will follow the same interview process as in Interview #1. However, the questions will focus more on how participants interact with the Citizen Helper interface and their thoughts on how they think the computer works to label tweets. The participants will be the same as the first round, so they will keep their participant numbers for this interview. The interview should take no longer than 1 hour.

B.2 Interview Start

[Script] Thank you so much for allowing me to sit in on your coding session today and ask you questions about your process.

Make sure to change Zoom name to participant number and press record.

B.3 Opening Prompts

As you begin the coding process, prompt them to focus on these two things:

[Script] As we go through this coding session, we want you to think about sharing details on these two things:

1. The computer interface that you're working with (Citizen Helper), and
2. Talk through what the computer would find easy and hard. (What would confuse the computer)

B.4 Tweet Labeling Task Questions

Ask these questions, when appropriate, during each interview session.

At this time, we will ask them to start their coding and speak out loud as they think through their coding decisions. These are the probing questions we will use to better understand their decision making.

1. Describe why you decided to code that tweet in this particular way.
2. Have you seen tweets like this before?
3. Was that a difficult tweet for you to label, and if so, why?
4. Was that an easy tweet to label and if so, why?
5. With this particular tweet, what do you think the computer might find hard?
6. What might the computer find easy?
7. Why do you think the computer gave you this tweet to code?
8. What is your understanding about why the computer selected this?

If you see them do a workaround (i.e., use the interface in a way not outlined in the training):

1. How did you figure that out?

B.5 Post-Session Questions

At the end of the session (at 45 minutes), ask these questions.

1. Think about using Citizen Helper. What did you find most frustrating in using it?
2. What do you like the most about using this system?
3. When you are struggling to decide how to code a tweet, what resources do you use to help you? (e.g., training documents)
4. What in your background do you believe helps you the most when coding these tweets?
5. Thinking about your background, how might that help the computer better learn how to automatically code tweets?
6. Is there anything I haven't asked you that you think might help me better understand how you worked today to accomplish your tasks.

APPENDIX C. PROTOCOL FOR INTERVIEW #3

C. TRAINING SCENARIO AND DECISION MAPPING/ILLUSTRATION

C.1 Overview

In the final interview, interviewers will follow the same process as noted in Interviews #1 and #2. However, there will be an emphasis on getting participants to share more in-depth details about their decision-making process. Interviewers will ask participants to imagine they are training the interviewer on how to label tweets and to draw out their decision-making process on a piece of paper.

C.2 Interview Start

[Script] Thank you so much for allowing me to sit in on your coding session today and ask you questions about your process.

Make sure to change Zoom name to participant number and press record.

C.3 Opening Prompts

[Script] As we go through the labeling session, we want you to label tweets normally, but on the more complex tweets, we will stop you after you have labeled them and ask you to more

thoroughly think through how you decided to label them. During this process, we will ask you to do one of these two things with the more complex tweets:

1. Imagine you needed to train me/us on how to label this tweet. Talk through how you would train me.
2. We are going to ask you to use a piece of paper and a pen to draw out how you think about labeling this tweet. This isn't an art project, and you can use boxes, stick people, anything that can illustrate how you are thinking about the labeling process. Once you finish your sketch, we will ask you to show it to us on the camera (or email it to us if you have no camera), and then you will explain your sketch. Our goal is to better understand how you think about the labeling process, so please don't worry about making it pretty.

C.4 Tweet Labeling Task Questions

Ask these, when appropriate, during each interview session.

At this time, we will ask them to start their coding and speak out loud as they think through their coding decisions. These are the probing questions we will use to better understand their decision making.

1. Describe why you decided to code that tweet in this particular way.
2. Have you seen tweets like this before?
3. Was that a difficult tweet for you to label, and if so, why?
4. Was that an easy tweet to label and if so, why?
5. With this particular tweet, what do you think the computer might find hard?
6. What might the computer find easy?
7. Why do you think the computer gave you this tweet to code?

8. What is your understanding about why the computer selected this?

Using the prompts above, ask participants the following questions on more complex tweets (do this 3-4 times at most).

1. Ask them the "train me" question on more complex tweets or
2. Ask them to draw out their decision process for more complex tweets

C.5 Post-Labeling Session Questions

At the end of the session (at 45 minutes), ask these questions:

1. What do you enjoy most about participating in the process of helping the machine learn?
2. What are some of the other tasks you have done as a CERT volunteer that you have enjoyed?
3. Tell me about your personal experience using Twitter. What about other social media?
4. May I ask you a few demographic questions before we wrap up?
    a. What type of computer/device did you use to label the tweets?
        i. PC
        i. Mac
        ii. iPad
        iii. iPhone
        iv. Other
    b. Age
    c. Gender
    d. Race/Ethnicity
    e. Years as a CERT volunteer