Social Reference Processing with Collaborative Human-AI Systems

by

Jordan S. Huffaker

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in the University of Michigan
2023

Doctoral Committee:

Professor Mark S. Ackerman, Chair
Professor Nikola Banovic
Professor Eric Gilbert
Professor Jonathan K. Kummerfeld

Jordan S. Huffaker

jhuffak@umich.edu

ORCID iD: 0000-0003-2876-9842

ACKNOWLEDGMENTS

Completing my PhD has been a long and difficult journey that I am pleased to say I have now finished. I have only been able to complete my degree due to the numerous forms of support I have received over the years, including from friends, family, and several organizations. I would like to highlight some of those that come to mind.

First, thank you to my committee: prof. Mark Ackerman, prof. Nikola Banovic, prof. Eric Gilbert, and prof. Jonathan Kummerfeld. You all have helped me tremendously in achieving the final milestones of my degree and you have given me confidence and inspiration in my research. Specifically, thank you Mark for persisting with me from the beginning, fresh out of undergrad. You have had a substantial influence on what has now been the majority of my adult life and you have shaped my worldview as a researcher. Also, thank you Jonathan for providing such valuable guidance for so many years. You have been a steady and persistent coach through the toughest parts of my degree, and you have helped foster much of the rigor in my dissertation.

Secondly, I would like to thank a few of my closest colleagues and friends in the computer science department who have together made up my strongest support network. Thank you Dr. John Chung for your steady interest in my research and your dedication to our friendship. Thank you for the many long chats we've had and for always lending an ear. Thank you Dr. Sai Gouravajhala for your enthusiasm and kindness, which gave me the motivation to push through the hardest parts of the PhD journey. Thank you Rebecca Krosnick for your persistent encouragement, which always fills me with confidence. Thank you Yasha Iravantchi for sharing such interesting ideas and for your incredible ability to troubleshoot the most random of problems. Thank you Nel Escher for your willingness to step into the nuances of my social research and provide valuable insights. Thank you Amani Alkayyali for spreading so much joy and for helping me find easy plants to take care of — even though I still managed to kill them.

Additionally, I would like to thank many more colleagues and collaborators throughout the University of Michigan community. These include those that have supported me from the Human Centered Computing lab, including Lei Zhang, Anthony Liu, Zhefan Ye, Divya Remesh, Jaylin Herskovitz, Dr. Harman Kaur, David Moon, Andrew Blinn, Sumit Asthana, Jane Im, Snehal Prabhudesai, Anindya Das Antar, Chen Liang, and Yaxuan Li. Thank you to the alumni prof. Sang Won Lee, prof. Jean Young Song, and prof. Yan Chen. Also, thank you to the faculty prof.

Alanson Sample, prof. Anhong Guo, prof. Xu Wang, and prof. Michael Nebeling who have each provided guidance at one point in my PhD. Thank you to my colleagues in the broader computer science department including Tianji Cong and Dr. Stephanie O'Keefe. Finally, thank you to Jeffrey Bilik who has been a friendly member of the broader UMich community and Hazem Turanni who has encouraged me to push myself in the UMich gym.

My PhD would not have been possible if not for the support and research experience I received as an undergrad at Purdue University. Thank you Alex Quinn for your mentorship during my undergrad. Starting out, I knew nothing about web development or research practices, and so I especially thank you for your willingness to walk me through the basics. My work and continued interest in crowd research was largely influenced by you. Also, thank you to my undergraduate lab members Dr. Gaoping Huang and Meng-Han Wu who both guided my initial research interests.

My work was financially supported by various organizations and a collaboration put together by prof. Jon Miller at the UM Institute for Social Research. Thank you prof. Jon Miller for the wisdom you've shared and allowing me to witness your research process. Thank you also to the other current and former members of the collaboration team prof. Belén Laspra, prof. Carmelo Polino, Dr. Glenn Branch, prof. Eugenie Scott, and prof. Robert Pennock. Thank you to NASA for their prior support of our collaboration, which partially funded my PhD. Additionally, thank you to Rackham Graduate School for covering various travel expenses and to Twitch.tv for their generous award in support of my research.

My collaborators at Lilt helped me expand the application of my research to a broader set of domains. Thank you Dr. John DeNero for your mentorship during both of my internships at Lilt and for your interesting insights. Thank you Geza Kovacs for your assistance in designing my experiments and for helping me overcome the barriers that prevented me from collecting data.

In addition to my research collaborators and friends at the University of Michigan, I have also been supported by many friends outside of the research world. Specifically, thank you Riley Bayless — my closest and longest-held friend — for being a constant rock in my life. I cannot put in words how influential you have been throughout my journey, not just while completing my PhD. Thank you also to Brooks Lobe for your reliable support throughout all aspects of my life. Thank you to my undergraduate friends who have remained interested in my journey, including Kyle Diekhoff, Eric Colter, Eric Nordin, Diego De La Fuente, and John Capella. I am grateful of my friends who have supported me since high school and whom I am still close with, including Josh Kirby, Dean Willis, Wyland Tregear, Teddy Pierce, and a newer friend, Braden McManus. Thank you to the Christensons including Victoria, Jason, Emily, Jim, Carl, Brittany, John, and Becky, as well as Kaitlin and Peter Morton (and many more) who have kindly lent their hospitality by housing me during Thanksgiving weekend and during various moves across the country.

Finally, I want to thank my family who have always been supportive of my research since the beginning. I am grateful to my brothers James, Jeremiah, and Justin Huffaker, and my sisters-in-law Paige, Madison, and Mel Huffaker. Thank you also to my extended family including my grandparents June and Bill Cook, and my aunt Justine Cook who helped me with my first study. I want to especially thank my late grandmother Alice Huffaker whose kindness will never leave my memory, and who generously donated the majority of the cooking supplies I use to this day. I continue to remember her every time I pick up the chef's knife with a hot pink handle she gave me so many years ago. Last but not least, I am so very thankful of my parents, Corinna and Jim Huffaker, who have always supported me in everything I do.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

ABSTRACT

In this dissertation, I introduce several human-AI systems that can understand and process *social references* — language that invokes connotations by overlaying parts of social and cultural contexts. I argue that developing a rich understanding of social references is a key challenge that must be overcome in order to accomplish numerous communication tasks, especially in online environments where language rapidly develops. I demonstrate the importance of understanding social references within three case studies: in finding emotionally manipulative language and distinguishing it from other aspects of content, in identifying and understanding rhetoric that is used to marginalize people-groups, and to ensure high-quality language translations for complex terms.

In each of these three case studies, I introduce *social reference processing* systems to overcome the challenges that limit existing systems, including systems that rely on the capabilities of crowds and machines on their own. Specifically, I use the case studies to introduce three collaborative human-AI approaches: a coordination technique that I call *anchor comparison*, a method for developing gold standards that can be used to identify social reference experts within a generalized crowd, and an approach for supporting people with gathering adequate context surrounding social references that I call *guided context building*. Together, each of my approaches make it possible to strategically recruit, coordinate, and support panels of crowdsourced experts with AI assistance for accomplishing effective social reference processing.

CHAPTER 1

Introduction

Modern social computing systems are now engulfed in numerous challenges such as containing the spread of misinformation, hate speech, and harassment. Despite the extensive funding and research efforts that go into building these systems, they are still extremely limited in their ability to wade into the complex social dynamics required to address many societal challenges. Namely, existing systems are generally limited to a find-and-delete paradigm for the most egregious content, where the most obvious cases are identified and removed. Solutions that provide more agency to users through personalized interventions would be more effective; however, crafting those interventions is still too computationally complex. As a result, existing systems can only marginally affect the spread of manipulative content and the growth of problematic trends like new othering campaigns.

I specifically examine the complexities around processing complex language. In particular, language thick with fast evolving social and cultural contexts escapes the capabilities of existing computation-only-based systems, but is commonly encountered online. At the same time, systems that require significant human labor to power are limited by the perspectives offered by the people that power them, which are often stretched thin by the scale of content they must attend to.

In this dissertation, I coin the term *social reference* to refer to terms that invoke connotations by overlapping parts of social and cultural contexts, and the term *social reference processing* to refer to a set of natural language processing problems that require knowledge and understanding of social references[1]. I argue that social reference processing problems are common in both commercial and academic settings, and that the unique challenges they pose have not yet been examined by existing work in the field. In my research, I bring many of these challenges to light by examining how social references are used within rhetoric and by introducing approaches for ensuring systems can identify and understand them.

Social references pose unique challenges to computational systems because they blend supplementary connotations together on top of more fundamental meanings. For example, consider the following segment from the former Fox News pundit Tucker Carlson describing congresswomen

[1] I will further discuss and define social references in the next chapter.

Ilhan Omar as a: "*living fire alarm*. A warning to the *rest of us* that we better change our immigration system *immediately*. Or else."[2] The phrase *living fire alarm* carries with it both the fundamental meaning of a literal fire alarm as well as the additional connotations surrounding the emergency of a fire. These additional connotations do significant work in the segment by inducing fear in audience members, making them think that that if they don't act now, they will be caught in the flames. After hearing the segment, one can almost feel the adrenaline from the phrase. Another example might include a phrase about the Jewish billionaire George Soros used in the following description of an immigration organization: "A pro-mass immigration organization with *links to billionaire George Soros* has successfully lobbied six Republican governors to resettle more refugees in their states." While the phrase 'links to billionaire George Soros' has the literal meaning of George Soros, who funds many organizations, it also carries with it many additional connotations that suggest a range of antisemitic conspiracy theories.

I find that existing systems used to process social references are either based in solely computation, the crowd, or a combination thereof. While computation-based solutions are desirable due to their low cost and are adequate for simple cases, systems that exclusively adopt them are inherently limited by their ability to understand social references that are captured within their training dataset. I argue that this limitation prohibits systems from operating without significant error in many online scenarios where language evolves quickly. Likewise, crowd-based solutions are not free of limitations: they are limited by the perspectives of the individual crowd workers that operate them. Individual workers may lack key context or may be prone to conflating social references with other content. Collaboration could solve this problem by including the perspectives of additional relevant people or information found externally, however, it is not yet known what expertise and context is necessary to understand social references, and how to obtain that expertise and context on-demand.

I argue that systems that leverage collaborations between both people and artificial intelligence, either through AI tooling or AI partners, are appropriate for social reference processing because they have the potential to combine the perspectives and expertise of groups of people with the consistency and scale of computation. I uncover the challenges that social reference processing tasks pose to these systems (that I refer to as *collaborative human-AI systems*) and I identify key insights that system designers should consider when using a similar approach. I additionally introduce novel approaches for coordinating, recruiting, and supporting groups of people with AI tooling, and I examine the implications of my approaches in three problem domains.

[2] From the July 9th, 2019 episode of *Tucker Carlson Tonight*: https://archive.org/details/FOXNEWSW_20190710_000000_Tucker_Carlson_Tonight/start/3540/end/3600

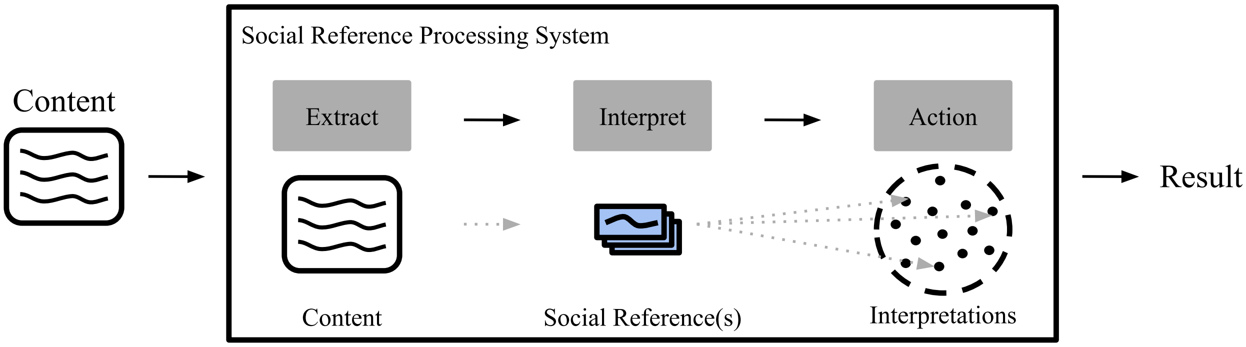


Figure 1.1: Social reference processing can be subdivided into three general subprocesses: 1) extracting social references from surrounding context, 2) interpreting social references into a space of possible interpretations, and 3) using those interpretations and an understanding of why differences may exist to take actions.

1.1 A Framework for Social Reference Processing

I make use of Ogden and Richards' idea of the *semiotic triangle* [169] to break down the social reference processing space into three main subprocesses: 1) extracting social references from content, 2) interpreting their meaning, and 3) using their meanings to take actions. I use these subprocesses as a framework to develop research questions and to understand the unique challenges associated with social reference processing. I will first explain the foundation of these three subprocesses and then I will explain how they fit together into a social reference processing system.

1.1.1 Social references as understood through the semiotic triangle

A social reference is a type of *symbol* that a person can understand by undergoing the interpretation process outlined in Ogden and Richards' semiotic triangle. In particular, a social reference does not carry an inherent meaning that is globally understood by all individuals, but instead a local one — each person must map a particular social reference into their own unique internal representation. People develop their internal representations through a socialization process.

Social references are a unique type of symbol in that they are commonly interpreted into a range of connotations (concepts) that often differ among people. These differences can be caused by a multitude of factors such as differences in people's internal representation of social references and differences in how people situate their understanding within particular contexts.

1.1.2 Social reference processing subprocesses

Instead of finding a singular meaning, social reference processing involves building a space of possible meanings and an understanding about why people's interpretations differ. I split this task into three main subprocesses (see Figure 1.1). The first of these subprocesses (that I call *extraction*) involves identifying social references and isolating them from surrounding context. This subprocess is made complicated since social references are often entangled with other kinds of language, making it difficult for both people and machines alike to understand where connotations are. The second subprocess, *interpretation*, involves creating a space of possible interpretations for specific social references. This subprocess is again made challenging due to the fact that people may substantially differ in their internal representation of social references. Finally, the third subprocess that I call *action* involves using the interpretations of specific social references and an understanding of why people's interpretations differ to make decisions.

As I will demonstrate throughout my dissertation, most social reference processing systems do not explicitly split these subprocesses apart, but instead do them implicitly as part of task-curated social reference processing workflows. For example, the vast majority of existing computational and crowd systems that are used to accomplish social reference processing tasks complete each of these three subprocesses in a single annotation step. Nonetheless, I found this division to be useful for conceptualizing the problem and for helping determine the direction of my research.

1.2 Research Questions

In this dissertation, I will answer the following main research question:

How can we effectively and reliably process social references?

Using the framework I have just described, I answer this main research question by exploring three additional research sub-questions:

RQ1: How can we disentangle the effect of social references from other factors within content by mitigating the biasing effect those content-factors might have on human annotators?

RQ2: What expertise is necessary to combine and reason with the context surrounding social references to understand their situated meaning, and how can we identify people with such expertise from a generalized crowd?

RQ3: What role can AI tooling play in ensuring that adequate context surrounding social references is available for human annotators to parse?

In my first research sub-question (RQ1), I sought to understand a primary challenge that makes it difficult for machines and people alike to complete the first social reference processing subprocess (extraction). I specifically explored the problem of identifying *emotionally manipulative* *language* (EML; e.g., using ‘living fire alarm’ to induce fear about immigrant congress members such as congresswoman Ilhan Omar) where I find that social references are a core delivery mechanism for invoking emotion in audience members and that they are difficult to identify due to how closely entangled they are with intrinsically emotional content (IEC; e.g., a parent losing a child).

Upon answering RQ1, it became clear that there are additional challenges that I would need to address to produce effective and reliable social reference processing systems for use in other domains. In particular, I learned the important role that context plays in completing the second social reference processing subprocess (interpretation), and that acquiring all of the relevant context needed to understand social references would be critical for expanding the capabilities of these systems. I coin the term *context building* to describe the process of gathering this relevant context and I explored two approaches for building systems that support the context building process.

In the first of these approaches, and in my second research sub-question (RQ2), I explored the idea of crowdsourcing an expert panel composed of people who have the relevant expertise and knowledge required to understand social references. I specifically wanted to expand the capabilities of my EML detection system by making it possible to detect *othering rhetoric* — rhetoric used to marginalize people-groups based on group identities (e.g., race, gender, disability, etc.). I sought to do this by learning about the expertise social justice experts use to identify othering rhetoric and by exploring how to find experts with relevant expertise.

My answer to RQ2 led me to explore another approach for supporting the context building process: with assistance from AI tooling. As I will discuss later in my dissertation, a key finding from answering RQ2 is that recruiting top experts onto an expert panel can be substantially difficult, but that recruiting panelists with adequate expertise is much more feasible. Upon this finding, I wanted to explore the possibility of boosting the capabilities of a panel by providing AI assistance.

I explored this alternative approach in my third research sub-question (RQ3), in which I examined the idea of using AI tooling to support the process of gathering and evaluating possibly relevant context surround social references. To achieve my end-goal of making it possible to apply this type approach to boost the performance of panelists in detecting othering rhetoric, I turned to another domain where human-AI collaborations are already commonly used to support the context building process: language translation. As I will discuss later in my dissertation, building context around social references is a core part of a professional translator’s duty, and that process is heavily influenced by AI. Part of the goal with RQ3 was to learn how AI tooling can better support this process for improving the efficiency and quality of language translations, which I could then pull lessons from and apply in other social reference processing domains.

1.3 Research Contributions

This dissertation introduces three contributions that make it possible to design effective and reliable collaborative human-AI systems for social reference processing. I assessed each of these contributions within three domains, I drew insights from each and I identified how they can be applied broadly to any collaborative human-AI system designed to process social references. My contributions are as follows:

1. **(RQ1)** In my first study, I explored the problem of identifying *emotionally manipulative language* (EML) where I uncovered the need to understand social references and differentiate them from other kinds aspects content. To explore the problem, I introduced the idea of *conflation error* which describes a kind of error human annotators make when they falsely annotate intrinsically emotional content (IEC) as EML. I introduced a collaborative approach for disentangling EML from IEC (called *anchor comparison*) that paraphrases the original content to remove EML, and then uses the paraphrased "anchor text" as a comparison point to judge the original content. I argue that my approach is a first step toward building systems that can identify social references. See chapter 3 for details.

2. **(RQ2)** Building on my first system, I conducted a second study that aims to extend EML detection to *othering rhetoric* and to uncover the processes and context people use to identify social references. In this second study, I examined the role social references play in creating rhetoric that targets and serves to marginalize people-groups based on group identities (e.g., race, gender, disability, etc.). My study aimed to make it possible to design a system that can recruit a panel of social justice experts from a generalized crowd that I call *Justice Panels*. I conducted interviews with experts in social justice where I learned about the skills, knowledge, and experiences participants use to find social references. From my data, I came to three conclusions that should be considered when designing systems that identify and recruit potential panelists. Most importantly of these is that social justice expertise is multifaceted, contextual, and situated, making it difficult to identify relevant expertise when needed. However, my study also contributes an optimistic finding: a promising gold standard that could be used to find people with relevant lived experiences for social justice-related tasks and a method for creating gold standards that could be used for many social reference processing tasks. See chapter 4 for details.

3. **(RQ3)** In my third study, I developed tools to help the process of uncovering and understanding social references. In this study, I introduced an alternative way of gathering relevant context in another domain that involves processing implicit connotations: language translation. Unlike my prior two approaches that rely on finding crowd workers with relevant

prior experiences to understand social references, my third system makes use of AI tooling to support the context building process. I examined AI approaches for supporting this process through a study of professional freelance translators (PFTs) where I cataloged how they make use of web resources such as concordances, dictionaries, and forums to build context around terms and their potential translations. I demonstrated that machine translation (MT) technology has a strong influence on how PFTs undergo context building, namely, by focusing their effort on checking suggestions provided by the MT. I introduced a new technique for facilitating collaboration between groups of people and machines for context building that I call *guided context building* that is capable of directing context building effort toward connotations where it has the highest impact. See chapter 5 for details.

1.4 Dissertation Outline

The rest of this dissertation will proceed as follows:

- Chapter 2 lays the foundations for social reference processing and reviews the relevant background literature.

- Chapter 3 describes my anchor comparison approach for disentangling social references from content.

- Chapter 4 describes my study of the expertise necessary to understand social references and of ways for identifying people with relevant expertise.

- Chapter 5 describes guided context building approach for assisting people with gathering relevant context around social references.

- Chapter 6 summarises the thesis, reflects on key findings, discusses limitations, and proposes future directions.

CHAPTER 2

Social Reference Processing

In this chapter, I provide background and motivation for building collaborative human-AI systems to accomplish social reference processing tasks. My work was inspired by scholars from a wide range of fields including social science, psychology, linguistics, human-computer interaction, natural language processing, political science, and other fields. I will provide a brief summary of that work in the following sections. First, I will discuss the foundations of my research in social reference processing, then I will discuss computation-based approaches NLP and HCI domains, then I will cover the challenges encountered with crowd-based and human-AI approaches.

2.1 Foundation for Social Reference Processing

I situate social reference processing as a sub-domain of natural language processing that involves processing particularly complex language. I formalize the concept of a social reference by pulling concepts from both semiotics and its sub-domain, conceptual metaphor theory.

Specifically, I define a social reference as terms that have the following two components:

1. A trope[1] that invokes connotations.
2. Connotations that require situational, discourse, and conceptual-cognitive context — either local or global — to decode.

In the first of these components, the term must use a literary trope to create one or more connotations in addition to a more formalized denotative meaning. A useful comparison would be to think of social references like musical overtones, they are connotations that implicitly accompany the canonical meaning. In the second component, I make use of Kövecses's concept of situational, discourse, and conceptual-cognitive context [120] to define the types of context social references use to create meaning. In his catalog of contextual factors that influence metaphorical conceptualization, Kövecses refers to situational context to describe the physical environment, social,

[1] Here I refer to *literary tropes*, as in metaphor, synecdoche, irony, etc. [32]

and cultural situation that influences one's understanding of a metaphor. This includes a person's knowledge of a particular community and their understanding of that community's characteristic interests and concerns. Likewise, Kövecses uses discourse context to encompass prior knowledge of the main elements of relevant discourse, in addition to broader discourse such as those that make up a person's ideology. Finally, conceptual-cognitive context refers to the conceptual system that is required to synthesize meaning from multiple concepts into a unique combined meaning.

My concept of a social reference is related to several other concepts in various literatures such as multivocality [170], polyvalent performance [205], and dog-whistling [80]. Social references can take on the form of a dog-whistle, which as Goodin and Saward describe, are *a way of sending a message to certain potential supporters in such a way as to make it inaudible to others whom it might alienate or deniable for still others who would find any explicit appeal along those lines offensive* [80]. Similarly, the effects of social references on the audience can be multivocal: as Padgett and Ansell explain that *single actions can be interpreted coherently from multiple perspectives* [170]. Albertson provides an example in a political setting where they demonstrate how George Bush used religious dog-whistles to create a multivocal appeal such that an in-group audience interprets the religious connotations but the religiously diverse out-group does not [6]. Social references can also play a key role in an actor's polyvalent performance, as Tilly defines as *individual or collective presentation of gestures simultaneously to two or more audiences in a way that code differently within the audiences*. My term, *social reference*, is inspired by these concepts in that it refers to terms that are rendered invisible to machines due to a lack of social and cultural context, but has meaning to people who do have that context available.

Social references play a key role in contributing to the overall rhetoric espoused by a particular actor. In particular, social references are important for conveying stereotypes since they can play the role of a semiotic payload to signal social groups, values, or beliefs [43]. An example might include using the phrase "dirty crime-ridden cities" which invokes a set of racist beliefs for many Americans by referencing a stereotype about minority groups living dirty lifestyles. Similarly, social references can have conspiracy overtones, such as mentioning that a particular immigration program has "links billionaire George Soros", which signals a range of anti-semitic conspiracy theories. They can also be used within stories and connected together to form narratives [93]. For these reasons, I argue that the ability to process social references is key to unlocking deeper meaning beyond that of which can be trivially captured within a dataset.

2.2 Computational approaches

Next, I cover a sample of computational approaches for working with language that includes social references. While no existing approaches are geared towards understanding social references outright, I examine several domains where social references commonly exist, and I discuss the shortcomings of these approaches in their respective domains.

2.2.1 Sentiment and emotion

While there are no natural language processing approaches for identifying social references explicitly, there has been extensive work on related classification tasks, such as sentiment analysis and emotion detection. Sentiment and emotion are a key part of rhetoric, and can be difficult to separate from the intrinsic content of text, similarly to distinguishing social references from content like I consider in chapter 3. However, much of sentiment analysis research has focused on product or movie reviews [136, 192, 225]. It has also been used to examine organizational communication, teamwork, and marketing — all in the social media, email, or other communication mechanisms [52, 73, 176]. More recent work has broadened the types of texts considered to include a range of topics on Twitter, but all labeling is completed by simply asking crowd workers to judge sentiment on a scale [49, 182]. Work on emotion detection has considered a wider range of text sources, but methods rely on expert annotated data that is collected with heuristic approaches to focus on examples expressing emotion [37].

Some of the most widely cited work in sentiment and emotion classification is the creation of lexicons specifying words with particular emotional content [64, 199]. While effective for the tasks mentioned above, these resources fall short of identifying sources of manipulative rhetoric that I consider here, in which words are used creatively, exploiting references to recent events, dog whistles and more. Prior work has explored bootstrapping existing context-based approaches by using knowledge bases or video footage in order to evaluate the meaning of these references [38, 197], but such solutions do not scale to provide knowledge of the history, culture, and community norms necessary to evaluate references that are grounded in social and cultural contexts — like many of the emotive phrases seen in news articles [70]. Systems that do not rely on such lexicons instead rely on large collections of manually annotated data. These are expensive to create with experts and crowdsourcing faces difficulties that I discuss in the next section.

2.2.2 Hate speech, harassment, toxicity, and related concepts

Likewise, automated approaches for identifying concepts such as hate speech, harassment, and toxicity struggle to grasp novel uses of language such as new social references. Approaches for this classification vary widely, as can be seen in Fotuna and Nune's survey of the field of automatic hate speech detection [68]. They found that existing work uses a wide range of text mining approaches, including N-grams, parts-of-speech, rule-based approaches, sentiment analysis, and deep learning. Their survey also identified some existing work that builds hate speech features by examining stereotypes and othering language (language that treats in-groups with superiority). However, they acknowledged several challenges that make automated approaches struggle, such as low agreement in hate speech classification by people, expertise that is necessary to understand cultural and social structures, and rapidly evolving environments that could change the meaning of language.

Burnap and Williams made use of a more advanced approach to hate speech detection that involves capturing typed dependencies (the relationships between words) that contribute to othering language [33]. They found that this approach works better in cases where more than one attribute of a protected group is being attacked. Alorainy et al. built on this idea by making use of intergroup threat theory to develop an othering feature set [8]. They then used embedded learning to determine parts of speech that are considered to be a part of an othering narrative. They showed that their approach outperforms existing techniques for identifying othering language, such as making use of lexicons, bag-of-words, and probabilistic language parsing approaches when tested on unseen datasets. However, both of these approaches focus on the language that is used to distinguish between social groups and the relationships between those groups. Uncovering the broader narratives as induced by social references that exist within text and their relationship to the overall othering rhetoric is still an open challenge.

Wulczyn et al. explored a similar concept by training a model to identify personal attacks as seen within comments [216]. They used a crowd approach to build a dataset, which they then used to train a classifier. They used the classifier to learn how personal attacks affect online discussions. They found, for example, that personal attacks make users much more likely to block the offending user. Breitfeller et al. similarly trained a classifier that works in-the-wild, specifically for microaggressions [28]. They made use of Sue et al.'s typology of microaggressions [201] to categorize the microaggressions that occurred within two online datasets, including a self-reported one and a crowdsourced one, and then they used those datasets to train a classifier. Like in prior studies, they arrived at low agreement scores between crowdsourced annotators, indicating their approach struggles with ambiguous language.

More recent research has expanded hate speech detection into image-based and implicit domains, but has continued to find limitations. For example, Aggarwal et al.'s analysis of multimodal models designed to detect hateful memes found that these systems are vulnerable to adversarial attacks where even very simple adjustments to input images can render hateful images invisible to models [2]. Mnassri et al.'s investigation into jointly training models to detect hate speech with emotional features found they can slightly reduce false positive errors over a baseline model, but that these improvements were not significant when used for offensive language detection [156]. Likewise, Huang et al. explored how ChatGPT could be used to detect and generate natural language explanations for implicit hate speech [101]. They find that its explanations are often better than human-written ones, but that they can be deceptively persuasive in cases where the model makes the wrong classification. Finally, Ghosh et al. developed a system for detecting implicit cases of hate speech in online conversation trees by making use of a user- and conversational-context synergized network. Their approach jointly models the conversational context and the author's historical and social context in the hyperbolic space to make classifications; however, their approach is still limited to cases that do not require world knowledge, such as the physical situation and social and cultural environment that may influence the meaning of the content [74].

2.3 Crowd and human-AI approaches

Crowdsourcing and human-AI approaches have the potential to bolster the ability of computational approaches. I will cover a sample of some of the general approaches, and then I will discuss what I consider to be an important challenge of managing disagreements between crowd respondents.

2.3.1 General approaches

Crowdsourcing is a common approach for overcoming the limitations of automatic approaches, often applied to gather training data [109, 191] or to integrate human intelligence into computational processes [20, 175]. Again, there are significant gaps in the crowd's ability to disentangle social references. Traditional approaches to achieving quality responses are to improve task instructions [137, 215], train or screen workers [153], or to decompose the task into subtasks [121]. Newer approaches leverage the crowd's reasoning skills to improve results [41, 60]. While these approaches are useful for addressing challenges like task misunderstanding, low quality workers, and groupthink, they were not designed to help workers recognize social references themselves — which I show in chapter 3 is still hard for even high-performing workers who understand the instructions clearly. Decomposing the task might make more sense, but there is no obvious way to decompose subjective judgment tasks for recognizing social references yet.

Alternatively, a line of work has explored the crowd's ability to make subjective judgments and mitigate bias. The crowd is particularly well suited for making subjective judgments because they can leverage social and cultural knowledge to predict how others might answer the same task [44]. Prior work has leveraged the crowd's ability to make these judgments to build emotion lexicons [158], assess image quality [177], and curate content based on personalized preferences [149], among other applications. Additionally, there is a line of work that attempts to mitigate biases in crowd worker responses. Prior work has explored bias mitigation strategies including those that warn workers of potential biases [103, 104, 185] and those that leverage statistical methods to debias results after-the-fact [107]. While this work helped me form an understanding of the problem, I note that disentanglement differs from bias mitigation in that specific biases cannot be known ahead of time. Without knowing biases, intervention-based approaches cannot be easily applied, as they would result in vague warnings to workers. My work aims to close these gaps by enabling new crowd collaborations that can parse social references.

2.3.2 Disagreement

While crowdsourcing and human-AI approaches do not suffer from the same limitations as automatic approaches, they do come with the challenge of managing disagreement among human annotators. As mentioned before, these approaches are attractive to both platforms and system designers alike because they can provide human insight at the speed and scale at which viral content emerges. Several forms of crowd-based approaches have been tried in the literature. Many of these use crowds to build labeled datasets that are later used to train machine learning models. Examples include for hate speech detection [194], toxicity detection [12], and credibility [153], among other related concepts. Aroyo et al. described two prototypical examples where a dataset is either constructed by averaging the absolute ratings of a few crowd workers, or by inferring ratings from comparative judgements [12].

All of these crowd-based approaches work by aggregating together the responses of many crowd members into a single "ground truth". However, individual members of the crowd may not agree with each other, and as a result of aggregation, their individual perspectives and lived experiences are merged into one. This could lead to a distorted perspective, where minority opinions are represented as the majority or where minority opinions are not represented at all. For this reason, Kairam and Heer argued that disagreements among crowd responses can actually indicate important information, not simply noise [112]. In their study, they introduced the idea of clustering together similar annotators to allow a systems designer to understand the perspectives represented within a dataset, or to potentially train multiple classifiers that can understand different perspectives. Gurari and Grauman's CrowdVerge system expanded on this idea by introducing an approach for predicting disagreements among crowd workers so that systems designers can gather adequate responses for dataset items that carry more disagreement and less for those where crowd workers agree [92]. Their approach is a more efficient way of gathering data from the crowd, however, it cannot yet uncover the source of disagreements.

An alternative approach to handling disagreement is through discussion. For example, Schaekermann et al. demonstrated that discussion groups helped eye care professionals understand the rationale behind diagnosis and helped them reach a consensus [186]. Chen et al. systematized this process in a crowd setting by coordinating crowd members to explicitly lay out their reasoning for their responses [40]. Part of the contribution of my study is that it provides insights into how a precursor system could be designed to complement systems such as Chen et al. by providing perspectives from people with meaningful lived experiences to inspire discussion.

It is important to uncover the reason annotator responses differ, particularly to ensure that the annotators are appropriately representative. Chen et al. contributed to solving this problem by introducing a mechanism that uncovers whether differences in annotator responses are due to genuine disagreements, or whether they are due to ambiguity [39]. In their approach, they asked respondents to annotate both the lower and upper bound of acceptable numeric ratings, which can then be used to parse both their own uncertainty and differences among respondents. Gorden et al.'s Jury Learning attempted to ensure appropriate representation in annotator responses by creating juries where the system designer can select characteristics they would like to be represented within members of the jury. Their system then trains classifiers to learn the prototypical perspectives of jury members, which are then used to generate classifications, and finally aggregated into a single response that includes important minority perspectives [82]. While these efforts take steps to understand sources of disagreement and ensure appropriate representation, they all still rely on cases where the designer knows ahead of time about the nature of possible disagreements. I help alleviate this challenge by identifying types of disagreement that a system designer should look out for, and by differentiating between disagreements and expertise.

I have discussed the literature that I used to lay the foundation for social references and social reference processing. In my review I have covered my definition of social references and I discussed common computational, crowd, and human-AI approaches a system designer might choose to use today. However, despite my search, I have found that there are still significant gaps between the abilities of existing systems and the needs required of social reference processing tasks. Specifically, there are three main gaps: 1) it is still not know how to differentiate between social references and surrounding context, 2) it is still not well understood what aspects are involved in social reference processing expertise, and 3) it is still not known how to support human teams with acquiring adequate relevant context to understand social references. For this reason, I have formulated three research questions in the previous chapter intended to fill these gaps in the literature. In the next chapter, I will discuss my exploration of my first research question.

CHAPTER 3

Anchor Comparison: Disentangling Social References From Content[1]

In this chapter, I examine the case of manipulative language where I demonstrate that social references are an integral delivery mechanism of manipulation, but are difficult to detect. I specifically consider one core challenge that leaves classification an open problem for both automatic and crowdsourcing approaches: distinguishing intrinsically emotional content (IEC; e.g., a parent losing a child) from emotionally manipulative language (EML; e.g., using fear-inducing language to spread anti-vaccine propaganda). Machine Learning approaches only work in narrow domains where labeled training data is available, and non-expert annotators tend to conflate IEC with EML. I introduce an approach, *anchor comparison*, that leverages workers' ability to identify and remove instances of EML in text to create a paraphrased "anchor text", which is then used as a comparison point to classify EML in the original content. I evaluate my approach with a dataset of news-style text snippets and show that precision and recall can be tuned for system builders' needs. My contribution is a crowdsourcing approach that enables the non-expert disentanglement of an important type of social reference from its surrounding context.

3.1 Introduction

Rhetoric that plays to people's emotions (e.g., fear-mongering rhetoric) can be an effective tool for inducing an emotional reaction in readers. Such reactions can cause "cognitive short-circuiting" [123, 129], resulting in the affected party taking actions or considering ideas they may otherwise disagree with (or even find repulsive). This is particularly true for emotions like fear and anger which increase susceptibility to tribalistic reasoning and inhibit empathy toward out-groups [89, 188]. Many actors exploit these effects for advertising [84, 143], increasing political influence [51], and amplifying misinformation, hate speech, and other harmful content [139]. I

[1] This work was previously published in [105].

explore approaches for flagging potentially-manipulative emotional language in text to facilitate future counter-measures such as de-ranking or nudging information seekers away from hate speech and misinformation.

Specifically, I explore crowdsourcing methods that can overcome the challenges inherent in separating *emotionally manipulative language* (EML) from *intrinsically emotional content* (IEC) — that is, separating dramatic presentation that is meant to stir emotion in the reader (EML) from content that may be emotional on its own (IEC). For example, IEC might include an account like Camila's: "Being an illegal immigrant means living in fear of deportation....My 19-year-old brother was deported when I was 17....It's been seven years now that I haven't seen him and don't know if I ever will."[2] On the contrary, EML induces emotion with language cues, such as in a Fox News segment where Tucker Carlson claimed congresswoman Ilhan Omar is a "*living fire alarm*. A warning to the *rest of us* that we better change our immigration system *immediately. Or else*."[3] Carlson used EML to play to people's xenophobia, reminding viewers that Omar immigrated from Somalia (an out-group), and signaling to them that "foreigners" are taking over the nation by pushing progressive policies.

I show that automated approaches and standard crowdsourcing approaches fail to adequately make this distinction. It would be preferable to solve this problem using automated approaches due to their low cost and scalability. However, while I expect such approaches to be capable of finding the simplest cases of EML, they are constrained by the labeled datasets available to them and they lack the ability to understand social references. For example, it is difficult to proactively include phrases like "*lispy queer*", "*living fire alarm*", and "*bad hombres*" in training data because of their creative nature, yet they carry salient cultural implications that can induce intense emotional reactions. While crowdsourcing approaches are more expensive, they are cheaper than hiring experts and can be leveraged at any point during the 24-7 cycle of the information ecosystem by recruiting workers from a workforce such as Amazon Mechanical Turk (AMT). However, I find that non-experts struggle to classify EML in text because they lack the ability to disentangle sources of emotion (IEC and EML). When asked to find EML, they tend to conflate IEC in their judgment (an error I call *conflation error*). Cognitive psychology offers a possible explanation for this result: people are inclined to substitute hard judgments with easier ones, such as substituting EML detection with their affective state [110]. There are currently no known crowdsourcing approaches for dealing with this problem. Approaches for mitigating bias in crowdsourcing settings come closest, but they can only offer general warnings that still leave workers with the challenge of disentangling sources of emotion.

[2] https://web.archive.org/web/20181210121052/https://www.manrepeller.com/2018/01/immigration-stories.html

[3] https://archive.org/details/FOXNEWSW_20190710_000000_Tucker_Carlson_Tonight/start/3540/end/3600

Instead, I propose a novel crowdsourcing approach I call *anchor comparison* that neutralizes the overpowering influence of IEC by measuring EML through comparison. I leverage workers' ability to identify and remove instances of language used to induce emotion in text to create a paraphrased "anchor text", then use that anchor as a comparison point with which to classify the original content. My approach prevents extraneous factors within content from influencing classification by constraining judgment to differences from an anchor, ensuring that only EML is reflected in measurement. More generally, my approach is the first that can disentangle social references from content.

Anchor comparison classifies text for whether it contains EML while giving systems builders the ability to tune precision and recall, a useful feature for accommodating different applications. In my motivating interaction, I envision a system that uses EML detection to identify potentially-manipulative content and warns the user. For this type of system, a false positive would represent content that is flagged for EML despite containing no EML and a false negative would represent content that has EML going unnoticed by the system. False positives of IEC (conflation error) would be particularly problematic in cases where controversial content is unfairly flagged, such as a post by a social justice advocate describing allegations of workplace harassment. I evaluate my approach by testing its ability to classify short text snippets as containing EML. I create a small dataset of short text snippets adopted from news articles, then systematically modify each snippet to create a version with heavy EML and one with very little EML while maintaining the same information between versions. I balance my dataset to include some stories with IEC and some without and measure classification performance with standard metrics (i.e., precision and recall).

In this chapter, I make the following contributions:

- I identify a class of problems that involve disentangling social references from content (e.g., EML and IEC). They are too challenging for non-expert human annotators, who have a tendency to conflate content with references (an error I call *conflation error*).

- I introduce an approach, *anchor comparison*, that transforms classification problems into a comparison task to mitigate conflation of content (e.g., IEC) and social references. I leverage workers' ability to identify and remove instances of language used to induce references in text to create a paraphrased "anchor text", then use that anchor as a comparison point to classify the original content.

- I build a system that leverages anchor comparison to distinguish between intrinsically emotional content and emotionally manipulative language.

- I evaluate my system on a small dataset of short text snippets adopted from news articles and demonstrate both the limitations of existing approaches (i.e., automatic and standard crowdsourcing approaches) and the feasibility of my approach.

3.2 Background

This work was motivated by a large set of research areas. In this section I synthesize the literature that guided my system design: 1) emotionally manipulative rhetoric and how it effects information processing and 2) the strategies media manipulators use to shape the information ecosystem.

3.2.1 Rhetoric, Emotion, and Reasoning

Emotion is an integral part of how people perceive, process, and leverage information [129]. The role of emotion as a persuasive tool has been examined by scholars dating back centuries, notably including Aristotle who argued that pathos (appealing to an audiences' emotions) is one of three pillars for effective rhetoric [11]. Despite being an effective tool, scholars agree that the use of emotion becomes problematic when it is used to overwhelm a reader's ability to think rationally [29, 138, 152]. Informally this is referred to as an emotional appeal, an argumentative fallacy that encourages poor reasoning [61]. While there are many ways one can manipulate a reader's emotions, in this paper I focus on the ones that use emotional words and phrases to do so. Content that manipulates reader's emotions by adding an opinionated slant or by carefully including emotionally laden facts I leave for future work.

3.2.1.1 Psychology of Emotion

Emotion impacts peoples' decision making by triggering their fast-processing cognitive system, one of two systems people use to process information [111]. Lerner et al. described this process and introduced the Emotional-Imbued Choice model [129]. Notably this model establishes that emotions shape the content and depth of thought people exert. For example, high-certainty emotions (e.g., anger) can lead to increased reliance on source credibility, leading to decreased attention to argument quality and higher usage of stereotypes and heuristics [23, 24, 204].

Importantly, manipulating emotions can impact how people behave because they are tied to *cognitive appraisal*, specific dimensions of cognitive state (i.e., how much attention I pay to the decision, how certain a person feels their actions will lead to a specific outcome, etc.) that lead to predictable decision outcomes [190]. For example, inducing fear and anger increases vulnerability to tribalistic reasoning and can lead information seekers to take impulsive actions [89, 188]. One relevant work found that varying the content of a news article in the wake of the 9/11 terrorist attacks from an anger-inducing framing (discussing how Arabs were allegedly celebrating the attacks) to a fear-inducing framing (discussing how more attacks are to come) led participants to perceive more or less risk in the world and prefer policies that were more or less harsh on potential violators [128]. These studies provide a useful framework for understanding why emotion affects decision making. In the next section, I will discuss how adversarial actors have abused this framework to influence their audiences.

3.2.2 Media Manipulation

The rapid growth of social media has led to extensive changes in the media ecosystem, leaving it vulnerable to manipulation by a variety of actors [139, 195]. These manipulators have learned to game platforms' recommendation algorithms such as those that determine news feeds [209] and videos [131]. Given that two thirds of Americans get their news from social media [144], most people are exposed to content planted by adversarial actors and must discern which ideas conveyed are reasonable and which are harmful. When people come across content, they view it through the lens of their social identity. Content that threatens, acknowledges, or confirms their identity can create an emotional reaction rather than a reasoned one [172], and it is this response that is most responsible for the spread of misinformation, hate speech, and harassment-inciting content [91, 140].

Actors have learned to maximize the virality of their content by playing to people's emotions. For example, Vosoughi et al. found that viral hoax tweets were more novel and evoked stronger feelings of surprise and disgust than non-hoax viral tweets [209]. Additionally, Song and Gruzd found that anti-vaccination content was much more popular and was more likely to be labeled under entertainment categories than pro-vaccination content [193].

Finally, actors often play to their audience by leveraging social appeals and cultural contexts. For example, Lewis found that a network of far-right YouTube channels gave credence to one another by maintaining close social ties through hosting one another in their videos [131]. While many of these YouTubers held contradicting beliefs, they mask their inconsistency by highlighting shared values, like their understanding of internet culture and their reactionary stance toward current events and ideas (e.g., feminism, social justice, and left-wing politics). Looking closer at many click-bait "fake news" stories, Marwick found that successful articles often connected to "deep stories", or the larger narratives readers often hold (e.g., that the rural states are neglected in favor of big cities) [140]. Now that I have described how actors use emotionally manipulative rhetoric to manipulate the information ecosystem, I will describe related work in mitigation strategies.

3.3 Related Work

In this section I will describe existing approaches for countering the effects of manipulative rhetoric and for setting up interventions at scale. I will focus on contextualizing my approach within the body of intervention strategies.

3.3.1 Mitigation Strategies

Prior work has explored two primary ways to mitigate manipulative rhetoric: 1) blocking or reducing the reach of explicitly-harmful content and 2) nudging people away from potentially-harmful content with warnings.

The first strategy has been widely adopted by social media platforms to limit the reach of blatant click-bait articles by removing them or by down-ranking them in search query results and recommendation algorithm suggestions [21, 35]. However, platforms have a variety of reasons to be hesitant to use this power including legal, financial, and political concerns [75]. Therefore, they typically reserve blocking and down-ranking for only the most extreme content [36].

The second approach, briefly adopted by Facebook [162], is to display a message to potential readers of an article that it is "disputed", "false", or something similar in order to dissuade people from believing it. In controlled lab settings, this approach has been shown to slightly reduce belief in the article [46]. However, people can become dependent on such labels after getting used to them, leading them to be more likely to accept un-flagged false articles as true than they would otherwise were there no flagging [171]. More effective interventions prime information seekers to think more carefully about the content they read [27] or warn them about specific strategies used to manipulate them [47].

Since these intervention approaches require manipulation efforts to be detected at scale to perform interventions for users of social media, many have sought to build automatic detectors. Detecting manipulative content can be done by identifying a variety of credibility indicators [224], including source [141] and content-based indicators [14, 155, 173, 189]. Often, credibility indicators are buried in context and require social and cultural knowledge to uncover them, making them challenging to identify. For these scenarios, prior work has demonstrated that collective behavior of information seekers can be a powerful proxy for uncovering credibility scores, including collective attention [154, 221]. This work demonstrates that combining time series and aggregate attention behaviors can be used to predict the credibility scores of tweets, a finding that might also extend to news article shares. While aggregate information seekers' behavior may be used to uncover other credibility signals, this strategy can only measure the effects content has on their information seeking, and would be unable to detect more nuanced signals that cause such behavior. Instead, I propose a proactive approach that involves looking for a particular credibility signal (EML) before the content has the chance to impact information seekers. In the next section, I explain EML more clearly and I introduce the dataset I used to evaluate my approach.

3.4 Problem

Before I describe potential classifiers to detect EML, I will define my problem more precisely, describe a small dataset I created to evaluate a variety of detection approaches, and explain how I measure performance. For the purposes of this paper, I treat EML detection as a classification problem and I specifically target extraneous language intended to induce an emotional reaction in the reader. I leave related tasks such as identifying specific EML words and phrases within a larger body of text and detecting biased reporting due to skewed facts for future work.

I envision two main applications that would benefit from the simple classifier as I have described: one that uses detection for intervention purposes (e.g., personalized nudges) and one that uses detection for content moderation purposes (e.g., de-ranking offending content or flagging content for site managers). I am particularly motivated by the first of these applications, as EML detection could be a useful backend for a system that performs inoculation interventions (e.g., [47]) or a system that points out specific EML words and phrases in the original content. A false negative would represent content that has EML remaining un-flagged, which in my scenarios, might result in lower user confidence for the overall system. On the other hand, a false positive would represent content that is incorrectly flagged for EML, which could lead to over prompting of the user or unfair punishment of content.

As a first step, I focus on classifying short text snippets (¡200 words) to ensure that workers can read and comprehend the text in the timespan of a microtask. Future work will extend to longer text such as full news articles, posts, or threads. I explore two factors that may influence classification:

1. *Emotionally Manipulative Language*: Whether there is highly emotive language that adds no informational value to the text. This is the factor I seek to detect.
2. *Intrinsically Emotional Content*: Whether the information conveyed is emotional itself, regardless of the language used to convey it. Information seekers frequently must parse emotional stories from manipulative language. The common news media trope “if it bleeds, it leads” hints at the prevalence of dramatic stories in media [220].

3.4.1 Dataset

I developed a dataset of twenty news-style text snippets adopted from 10 news articles[4]. For each news article, I created a shortened version that includes heavy EML and a version that includes

[4] Can be accessed online at: https://doi.org/10.7302/yhpy-e679

	¬IEC		**IEC**	
	¬EML	*EML*	*¬EML*	*EML*
# snippets	5	5	5	5

Table 3.1: My test dataset is balanced across four conditions.

very little EML, creating a pair. I created each pair so that both versions had the same information, with the only difference being EML. I picked five news articles that contained IEC and five that did not, making the dataset balanced with five snippets per condition (Table 3.1).

I evaluated the quality of my dataset by hiring a journalist and a member of the editorial staff for a nationally prominent news magazine to rate each text snippet on a 5-point Likert scale along three dimensions. First, I confirmed the main factor of my dataset by asking reviewers to rate "How much does the paragraph intentionally stir emotion in the reader?". Second, I confirmed my variation of IEC by asking raters to assess "How much emotion is intrinsic to the information conveyed in the paragraph?". Finally, I sought to confirm that some snippets are publishable in reputable news sources by asking raters to agree or disagree to: "If the facts were correct, this is something that would be publishable in a reputable news source."

I met with each rater separately and followed a three part procedure for each snippet in my dataset: 1) I asked the raters to rate the snippet on the three dimensions mentioned, 2) I asked the raters for the reasoning behind their answers to ensure that they understood the meaning of each question, 3) I allowed the raters to change their answer after discussion. I reached a high Fleiss' kappa score with the two raters and myself about which of the snippets in each pair used more EML (k = 0.87), which snippets had a "decent amount"[5] and which had more intrinsic emotion (k = 0.60)[6], and which of the text snippets in each pair were more likely to be publishable by a reputable news source (k = 0.72).

3.4.2 Measures

My main outcome measures are standard classification metrics (i.e., precision and recall), focusing specifically on conflation error (false positive rate of IEC). Additionally, I define two more metrics that I will use later in this paper to evaluate intermediate outcomes: *EML level* and *distortion*. EML level is coded on a 4-point scale and indicates the relative amount of EML in a text snippet with 4 being the maximum and 1 being the minimum. In each snippet pair, I define the snippet with EML

[5] This was the middle option on a 5-point Likert scale.

[6] One of the raters found some of the text snippets more personal (and thereby having more intrinsic emotion) than the other rater and I found. To ensure these differences would not affect my results, I ran my evaluation with the codes provided by the differing rater and found only marginal differences in the results of my baseline approaches.

to be level 4 and the one with no EML to be level 1. Additionally, I use the metric distortion to code whether a change to a text snippet has changed its informational content. I use this dataset and described measures to evaluate various classification approaches.

3.5 Baseline Study

In order to understand the limitations of existing approaches to classifying EML, I first explore five logical baselines and find that none achieve satisfactory performance. In particular, I explore four automated approaches and a simple crowdsourcing approach.

3.5.1 Baselines

At first glance, it may seem that automatic approaches (e.g., machine learning) can be trained to sufficiently classify EML due to their success in classifying text in similar domains (e.g., emotion detection and sentiment analysis). For this reason, I evaluate state-of-the-art performance for *sentiment analysis* and *emotion detection* by classifying snippets from my dataset.

Emotion detection is the task of classifying the intensity of typically 4-6 emotions induced by a piece of text. Existing approaches make use of an emotional lexicon to match words with the emotion they commonly evoke [157], then aggregate those emotions to determine the overall emotion evoked by the paragraph. Specifically I evaluate:

> *IBM Tone Analyzer* [4]: This classifier was trained on Twitter customer service data and uses a support-vector machine (SVM) to classify emotion on a scale from 0-1 for anger, disgust, fear, joy, and sadness.
>
> *EMPATH* [65]: EMPATH is a tool for coding lexical categories (such as emotion) in large-scale datasets (similar to LIWC). The tool “adapts” to new dataset domains by enabling users to seed additional lexical categories, and then by recruiting workers from AMT to find related terms in the dataset. I use EMPATH to code emotion words.

I also considered sentiment analysis, which is the task of classifying positive, negative, or neutral feelings expressed in text. While sentiment analysis is less related to EML detection than emotion detection, the datasets for sentiment analysis are larger than those for emotion detection, and sentiment analysis should still be capable of detecting large attitude-slants. I explore:

> *VADER* [106]: VADER was designed to generalize to a variety of domains by combining features from a sentiment lexicon with five rules. Its rules were created from a qualitative coding of tweets and its simplicity makes the model perform well without large-scale training data, unlike most other approaches.

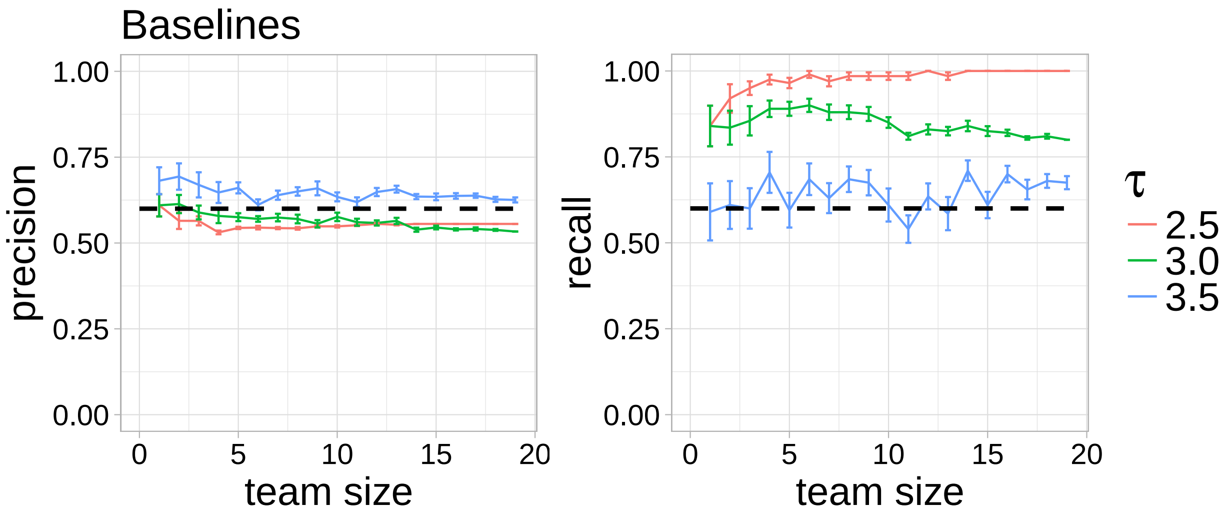


Figure 3.1: Precision and recall of the baselines. The dashed line indicates the best performance of my automatic baselines and "τ" indicates different decision boundaries on a 5-point Likert scale.

BERT with Fine-Tuning [56]: I trained a three-way classifier on top of the BERT-Base uncased pretrained model, using the 2017 SemEval sentiment analysis task data [182]. This approach takes into account the context of words by using a pretrained language model (BERT) achieving an F1 score of 0.636, slightly below the state-of-the-art for the dataset (F1 = 0.677)[7]. I used a Twitter dataset because it offered more diversity than the movie and customer review datasets, making it more likely to include information about current events and culture.

Finally, I consider a standard crowdsourcing approach for classifying EML that asked workers from Amazon Mechanical Turk (AMT) to rate text snippets on a 5-point Likert scale[8], takes the average of those ratings, then uses a decision boundary to determine positive or negative classification. For my evaluation, I recruited workers from AMT using LegionTools [81], presented them with a single text snippet from my corpus (requiring unique workers for every task), and asked them to answer "How much does the paragraph intentionally stir emotion in the reader?"[9].

3.5.2 Results

I find that the automated baselines only perform marginally better than a random baseline and the standard crowdsourcing approach performs slightly better than the automatic baselines for recall, but has similar precision (Figure 3.1). In particular, I find the IBM and BERT baselines have a high false positive rate while EMPATH and VADER had balanced error rates [10]. Additionally, I see that the crowdsourcing baseline had high conflation error (misclassifying IEC text snippets as having EML), explaining the low precision.

3.5.2.1 Automated Approaches

While it may be possible to improve the performance of machine learning approaches by training on a dataset explicitly labeled for EML, I argue that their performance will remain limited for three reasons: 1) generalizing knowledge to examples not explicitly trained for remains a challenge for even state-of-the-art models, 2) when these classifiers are deployed in the wild, adversarial actors will be highly motivated to find exploits, and 3) hiring experts to create labeled datasets is expensive and time consuming.

First, machine learning approaches are limited in their ability to generalize knowledge beyond what exists in training data, making them vulnerable to novel patterns and references not explicitly trained for. For example, phrases such as "*lispy queer*", "*living fire alarm*", and "*bad hombres*" are rare manipulative phrases that are unlikely to show up in training data, and so would likely go unnoticed. The use of cultural references are common, constantly changing, and can take many different forms, making it nearly impossible to include all of them in training data. As this limitation has become increasingly better known, scholars have started developing methods to test classifiers for their generalizability by developing "adversarial datasets", modified versions of existing datasets whose changes do not effect human performance but often tank machine performance [94, 165, 222]. I expect such datasets will be important for assessing future EML classifiers.

Second, prior work has found that adversarial actors are highly motivated to exploit algorithms that are deployed in the wild, often in a coordinated manner [79]. In particular, actors actively manipulate search engines to amplify their content by finding "data voids", search terms with limited data available to populate search results, then by posting extremist content that uses those

[7] I attempted to train the state-of-the-art model but experienced issues in the training process. Since my model performs similarly on the Twitter dataset, I believe that its results should be representative.

[8] Likert scale ratings are standard for related tasks [106].

[9] In a prestudy, I tried many ways to word this question in order to achieve better results. I found that wording only marginally effected results, and that my findings hold despite the choice.

[10] The IBM Tone Analyzer and EMPATH detectors output continuous scores. I convert these scores to classification categories based on whether they cross a decision boundary. For all decision boundary values, accuracy did not exceed 60% for either detector.

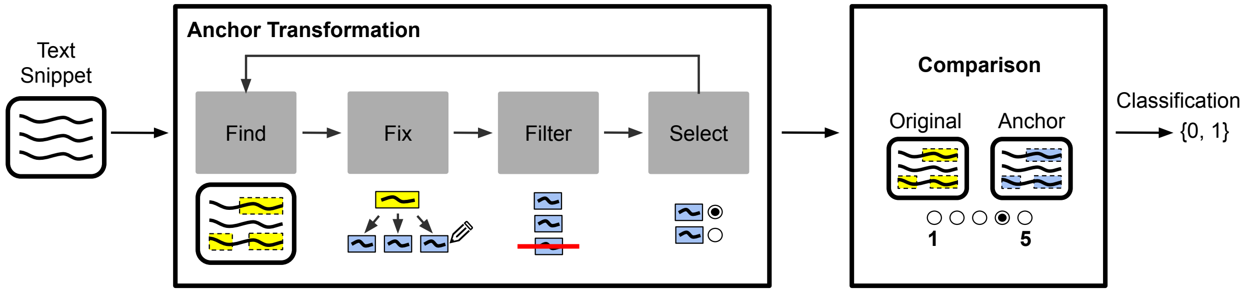


Figure 3.2: My system splits the task of classifying EML into two parts: anchor transformation and comparison. Anchor transformation involves removing EML from the original text snippet to form an “anchor”. This is done in four steps: finding portions of text that contain EML, suggesting possible edits to remove EML from each portion, filtering edits to remove those that introduce distortion, and selecting the best edit from a group.

search terms. For my envisioned interactions (i.e., content moderation and intervention systems), I expect that actors will also be motivated to find novel exploits, making it important that backend EML classifiers are reasonably robust against these attacks.

Finally, hiring experts to create a labeled dataset for EML detection is expensive and time consuming, making it infeasible to fix the previously described problems by creating massive annotated datasets. Since, as mentioned, cultural references are constantly in flux, over time new annotations would need to be gathered to keep up detection quality amid new current events and discourse themes. In addition to performing better, a crowdsourcing approach would be more cost effective than hiring experts, and thereby might be used to cheaply create a large-scale dataset for EML detection.

3.5.2.2 Standard Crowdsourcing

The simple crowdsourcing approach failed because workers tended to conflate IEC with EML. A chi-square test comparing worker ratings for snippets with IEC and no EML with their ratings for snippets with no IEC or EML confirmed that the tendency to conflate was significant $\chi^2(4)$ = 32.49, p ¡ 0.001 (effect size = 0.44).

This finding is supported by the theory of the affect heuristic in psychology, which contends that people commonly use their emotions as a cognitive shortcut to make judgments (i.e., how they are emotionally affected by the judgment affects judgment) [67]. While I cannot be certain that the affect heuristic is the only cause of conflation error in my context, it does offer an explanation for my results: workers are affected by IEC and are substituting the EML detection with their affective state by rating IEC highly. Unfortunately, this problem is not easily solved, as prior work has found that people struggle to disentangle sources of emotion even after being warned about the potential to attribute emotion to the wrong source [212].

Given that automatic approaches did not work sufficiently for this problem, and a standard crowdsourcing approach led to the challenge of disentangling sources of emotion, I developed a new crowdsourcing approach. In the next section, I will describe how I overcome the limitations of these baselines by transforming the judgment task into a comparison problem, thereby limiting the influence of IEC on final classification.

3.6 System

In this section, I will describe a system I created that leverages an approach I call *anchor comparison* to mitigate the overpowering influence of IEC on worker judgment. My approach decomposes the problem into two pieces (Figure 3.2): 1) anchor transformation which involves coordinating workers to remove specific instances of EML in the text to create an "anchor" version of the original and 2) comparison which involves measuring the difference in EML between the original and the anchor on a 5-point Likert scale. It thereby turns the classification problem into one of comparison, enabling a task that was previously considered to be atomic to be decomposed.

To explain why anchor comparison works, I will build off my previous psychological analysis of the problem. As I have previously explained, workers likely conflate IEC with EML because they use their affective state (how they feel) to decide how to classify the content [67]. IEC makes them feel strongly, which is then reflected in their Likert scale ratings. My approach works by anchoring workers' affective state in the anchor text and measuring only the difference from the anchor to the original text. Differences from the anchor may still affect workers, but this affect would be due to EML since IEC is held constant. My approach builds upon reference-based crowdsourcing approaches such as [218] in that I leverage the crowd to create the point of reference. In the next sections, I will walk through each component of my system. While I do so, I will describe a study I used to measure its performance and set parameters. I will conclude by describing two key tradeoffs my system affords.

3.6.1 Anchor Transformation

Anchor transformation is the process of transforming text into an "anchor", a paraphrased version of the original text that has been revised to remove EML. The problem shares many characteristics with copyediting with one key difference: particular sensitivity to "distortion" errors, where workers alter the information content in the text in an attempt to remove EML. Distortion can lead to false positives (and conflation error) in the comparison step by creating a fabricated difference between the original text and the anchor.

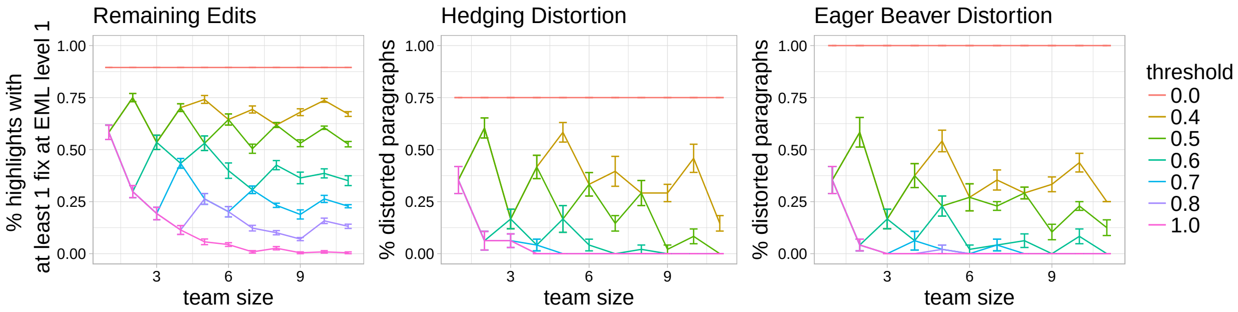


Figure 3.3: Results for the filter step. Increasing the agreement threshold reduces the number of paragraphs with a distorted suggestion (middle and right), but as a consequence, it also reduces the number of edits that make it through the system (left chart). For "eager beaver distortion" I allow one distortion since it only has a minor effect on the false positive rate, but setting the threshold high enough would eliminate even this error.

To enable explicit control of distortion, I build upon the current state-of-the-art for crowd-sourced copyediting (Bernstein et al.'s Soylent [20]) by adding a "filter" and "select" step. The resulting system consists of four steps: finding EML words and phrases, suggesting potential edits to remove EML from those phrases, filtering out suggested edits that have distortion, and selecting the best edits from those remaining. For each step, I hire workers to complete the task in parallel (hiding other worker's responses to avoid potential groupthink) and pay them at an hourly rate of $10 USD/hour. I will describe each step in detail below.

3.6.1.1 Find

The find step involves identifying parts of the text that have EML by enabling workers to highlight portions of the text, then aggregating based on highlight overlap. Specifically, I ask them to "Highlight dramatic[11] words and phrases in the [text]." In my component study I found that a 20% worker overlap threshold is the optimal value for maximal overlap with my annotations of the text and that performance plateaus at approximately 5 workers at 75% word-wise agreement.

Examining the portions highlighted by workers, I notice four types of highlights. The first are highlighted portions like "tragic deaths" and "sweet and unsuspecting American children" that can be fixed by simply removing the unnecessary verbiage (e.g., "tragic" and "sweet and unsuspecting"). Secondly, some highlights were made of entirely IEC like "Panama fungal disease threatens future crops" and require no editing. Thirdly, some highlights contain both EML and

[11] I switched to using the word "dramatic" instead of "emotional" after analyzing workers' qualitative explanations for their responses in a pre-study and noticing that many workers interpreted "emotional" in the instructions to mean "that I had an emotional reaction to" instead of my intended meaning "that the author was *trying* to get me to react emotionally to". I tried a variety of wordings and found that "dramatic" yielded the best results.

IEC like “Gary was swept by a wave of grief”. These highlights require clever rephrasing in order to maintain information while removing EML (e.g., rephrasing to simply “upset”). Finally, some highlights contain both EML and IEC, but are particularly challenging to rephrase in a way that maintains information while removing EML. These highlights would be better suited for a fix that restructures the sentence. Most highlighted portions landed in the third category, requiring clever rephrasing to remove EML while maintaining IEC.

3.6.1.2 Fix

In the second step I ask workers to suggest possible edits to remove EML from the highlighted portions in the previous step. I ask workers to provide an edit for all highlighted portions, scaling their pay based on the number of edits I ask them to make. Specifically, I provide the instructions: “Remove dramatic words and phrases from each of the highlighted portions while maintaining the same information and grammatical correctness.” To prevent workers from attempting to fix multiple highlighted portions in the same edit, I restrict the range of text that workers are allowed to edit to include text starting from the end of the preceding highlight (or beginning of the text if the first highlight) to the start of the next highlight (or end of the text if the last highlight). While I give workers the option to skip highlights that are too challenging to rephrase, they generally provide a suggestion for all highlighted portions anyway.

In my component study, I find that performance plateaus at about 5 workers. After fixing this parameter I observe that 75% of highlighted portions will have at least one suggested edit that correctly removes all EML from the portion.

3.6.1.3 Filter

My third step is to filter out suggested edits that distort the information of the original text. As I have noted above, distortion error can cause false positives later in the pipeline as workers conflate the change in information to be a change in the amount of EML. For IEC text, increased distortion can lead to increased conflation error.

Therefore, I build this component to include an *agreement threshold* that can be used to control the level of confidence that suggested edits passing through the system are not distorted. I ask workers to select suggested edits that “maintain the same information as the [original text]” and to “not select [suggested edits] that attempt to debias or soften the opinions in the [text]”, then I aggregate suggestions based on the percentage agreement between workers. Setting the agreement threshold higher allows fewer edits through, but at higher confidence they are not distorted. Setting it lower may allow a higher percentage through, but without as much confidence in their quality.

In my study of this component, I noticed two ways that workers distort information in the original text: 1) as a result of workers' attempts to soften the opinions in the text by hedging the claims (I call this *hedging distortion*) and 2) as a result of workers too eagerly removing inconsequential information in addition to EML (I call this *eager beaver distortion*). I notice that hedging distortion can have a large effect on downstream classification, but eager beaver distortion less so (Figure 3.5). Despite this, I observe that nearly all hedging distortion and most eager beaver distortion can be eliminated with a 50% agreement threshold at 5 workers while ensuring that 50% of highlighted portions have a high quality suggestion (Figure 3.3). Setting the threshold to 80% eliminates all distortion while ensuring that 25% of highlighted portions have a high quality edit.

3.6.1.4 Select

The final step involves selecting the suggested edits that best remove EML from each highlighted portion. Of the suggested edits that make it through the filter in the previous step, I ask workers to "determine how well [each suggested edit] removes dramatic words and phrases from the highlighted portion of the [text] while still making sense" on a 3-point scale labeled "best", "decent", and "worst". For each highlighted portion, I select the suggestion that has the lowest mean score and incorporate the change into the final paragraph. My component study found that 5 workers is sufficient for consistently selecting the best edits out of a group.

3.6.1.5 Iteration

I find that anchors can be iteratively improved by sending the output of the "Select" step back into "Find" step, reducing the EML level with each iteration. In my component study I found that iterations reduced the EML level of the anchor by an average of 1 point (of 4) with each iteration, eventually converging to an EML level of 1 after 3 iterations.

3.6.2 Comparison

After an anchor has been created in the anchor transformation step, I ask workers to assess the difference in EML between the anchor and the original text. Part of the contribution of my approach is that this step enables for the high-level task to be decomposed into highlight-level units. The comparison step aggregates these individual decisions into a final classification decision. I show workers the two versions of the text side-by-side and ask them to rank on a 5-point Likert scale "How much more dramatic is the [original text] compared to the [anchor]?" I aggregate ratings by taking the mean value and then by using a decision boundary to determine whether the text is to be classified as a positive or negative example.

My evaluation of this component found that classification accuracy is dependent on the quality of the anchor used for comparison (Figure 3.4). For an anchor that correctly removes all EML from

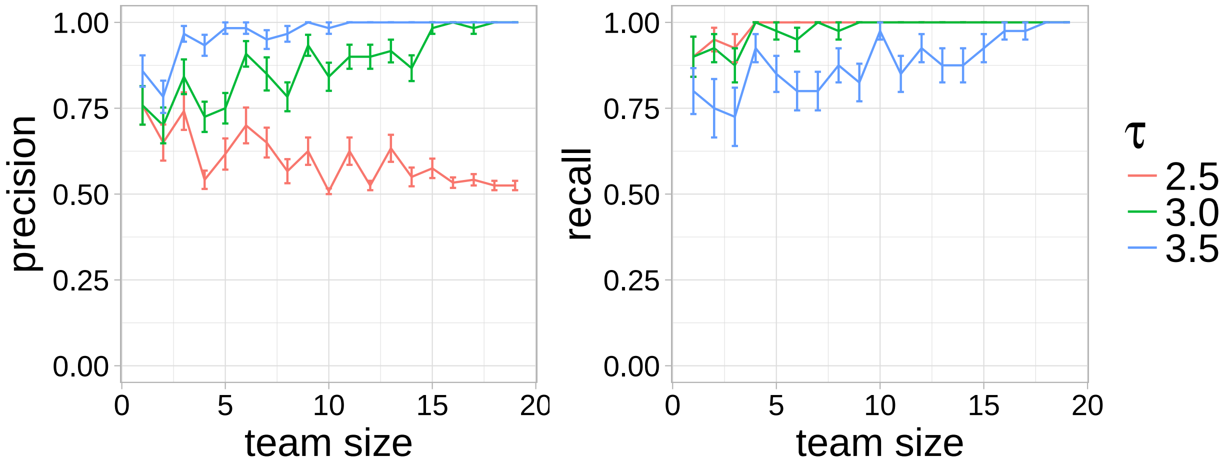


Figure 3.4: Precision and recall for the comparison step. "τ" indicates different decision boundaries on a 5-point Likert scale.

the text, 90% accuracy can be achieved with 5 workers and using more leads to even higher accuracy. However, I find that imperfect anchors cause classification error (Figure 3.5). For example, anchors that fail to completely remove EML (EML level greater than 1) are more likely to cause false negative classification because the difference in EML is perceived to be lower. Likewise, distortion in the anchor can create false positive classification because changes in information can be perceived as a difference in EML.

3.6.3 Putting It Together

Through a series of component studies, I have demonstrated the feasibility of anchor comparison. I showed that the anchor transformation process (left box in Figure 3.2) can create a version of the text with low EML and no distortion when a high agreement threshold and iteration are used. I also showed that the comparison process (right box in Figure 3.2) can achieve perfect precision and recall given the original text and a low EML version without distortion.

I will now walk through a version of my system that can achieve reasonable precision and recall. Given a set of new snippets for classification, I would first send them through the find step. On each iteration through anchor transformation (left box in Figure 3.2), I can first use 5 workers with a 20% overlap threshold, to highlight 75% of the EML words. Secondly, the snippets would go through the fix step. Using 5 workers, I can get high-quality suggestions for 75% of the highlighted portions. Thirdly, I send the snippets to the filter step. My data shows that I can use 10 workers with a 50% agreement threshold to ensure 100% of paragraphs have no hedging error and

87% of paragraphs have at most one eager beaver error. However, I note that the line trends down (right graph in Figure 3.3), indicating that including more workers would feasibly ensure 100% of paragraphs have at most one eager beaver error. While distortion trends down, I note that the % of highlighted portions with a high quality suggestion remains constant (left graph in Figure 3.3). With three iterations through these steps (Find, Fix, Filter, Select), I would reach a final version. Finally, I would send my snippets through the compare step. Given that there is at most one eager beaver error in each snippet, I would be able to ensure perfect recall and 87% precision. System builders can also save costs by tuning classification errors for their needs.

While cost depends on the number of edits needed to remove EML, my data had an average of 5 highlighted portions per iteration, leading to a projected cost of about $47 per snippet. While this may be feasible for some applications, future work will need to address the engineering of reducing costs.

3.6.4 Tradeoffs

My system affords builders two mechanisms to tune precision and recall. The first mechanism is *iteration* which can be used to control recall. Secondly, the *agreement threshold* can be used to control precision.

3.6.4.1 Iteration improves recall, but is more costly

The higher the EML level in the anchor text, the less the perceived difference between the original text and the anchor, and the more likely text with EML will be classified as having no EML. Iterations can be used to reduce the average EML level of anchors, and thereby reduce the false negative rate. However, increasing the number of iterations increases cost as more labor is required to find, fix, filter, and select edits.

3.6.4.2 Increasing the agreement threshold improves precision, but reduces efficiency

Allowing distortion in the anchor introduces the possibility that the distortion will be perceived as a difference in EML during the comparison step, leading to text with no EML to be classified as having EML. Increasing the agreement threshold can reduce the % of paragraphs with distortion error and reduce the false positive rate. However, increasing the agreement threshold has the side effect of reducing the number of suggestions that make it through the system, leading to a lower efficiency at which EML is extracted and causing an increased number of iterations needed to keep recall constant.

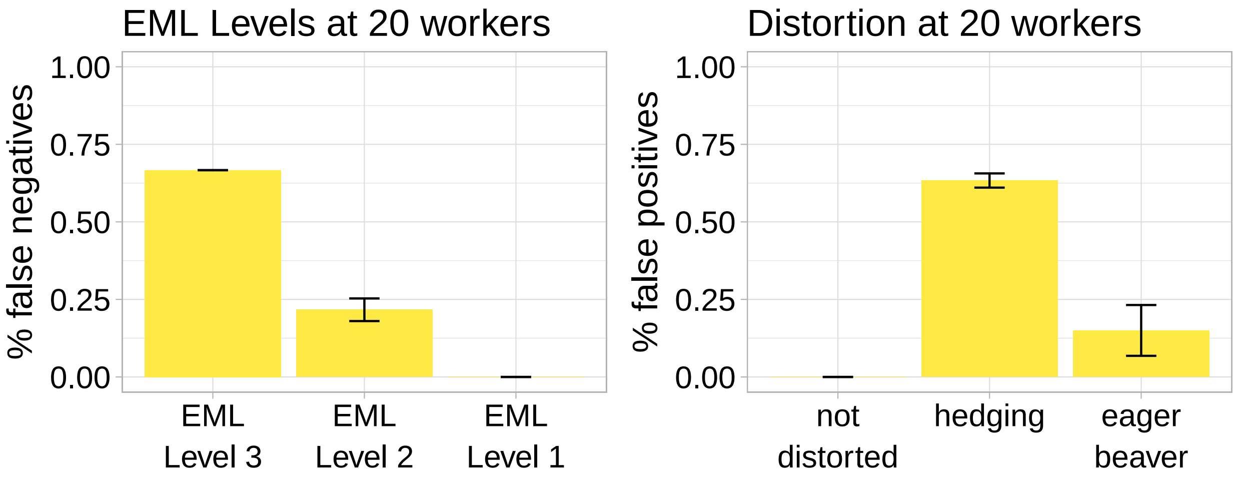


Figure 3.5: Errors in the anchor transformation step can compound into classification errors in the comparison step. Higher EML levels in the anchor cause an increased false negative rate and adding distortion leads to an increased false positive rate.

3.7 Discussion

In this work, I examined the challenge of classifying text for social references. A specific challenge that makes that classification difficult is that references are often entangled with content, which leads to the potential for one to be misconstrued for the other. I have proposed a new crowdsourcing approach that controls for conflation by transforming the classification problem into a comparison.

I then demonstrated the feasibility of my approach by exploring an important sub-problem that involves disentangling social references evoked through emotionally manipulative language (EML) from intrinsically emotional content (IEC). To test the limitations of existing approaches and the feasibility of my approach, I developed an appropriate test dataset and used it to evaluate five baseline approaches as well as ours.

I showed that existing approaches for classification struggle to distinguish between EML and IEC. Automatic approaches require substantial training data to understand social references which is both expensive to obtain and still leaves them vulnerable to novel language patterns that have not yet been added to training data. Crowdsourcing approaches perform better, but are prone to conflation error because non-expert human annotators struggle to disentangle sources of emotion in text — whether it is EML or IEC.

My approach, anchor comparison, overcomes these challenges by leveraging workers' ability to find specific instances of EML in text and draft edits that remove them, resulting in an "anchor text" that can be used as a point with which to classify the original content by comparing the two.

Through a series of component studies, I demonstrated that this approach is feasible and that it affords two mechanisms systems builders can use to improve precision and recall. At the penalty of increasing the cost of classification, iteration can be used to improve recall and an agreement threshold to improve precision.

My contributions are then, a class of problems that involve disentangling social references from content, my anchor comparison approach that leverages transformation for disentangling these references, a system I created that uses anchor comparison, and an evaluation of my system.

Limitations. However, my work has several limitations that remain for future work. First, the social references I address here assume that multivalent messaging is received by a homogeneous audience. Violations of this assumption (e.g., dog-whistles) have references for only sub-audiences and may need a specialized crowd to detect.

Second, I demonstrated through a proof-of-concept that anchor comparison can be used to detect EML without conflation error at a cost of about $50 given a short text snippet and sufficient run-time. In principle, this approach could be used to detect emotionally manipulative language online content. However, wide-scale deployment will require reducing cost. This remains for future work but I believe that this will be possible by combining machine learning approaches with the crowd. For example, one could use my crowdsourcing approach to generate a labeled dataset, which could in turn be used to train a classifier. Additionally, anchor transformation is particularly suitable for a hybrid intelligence approach, where machine learning is used to find and suggest potential edits to remove EML from the text, and crowd workers are used to make final judgments.

Third, while I took steps to ensure validity of my test dataset, my current implementation requires text snippets to be short and have all necessary context included in each snippet. Data encountered in the wild will likely include visual aspects, vary in length, and interconnected contexts. Future work will consider summarizing long news articles etc. that include all necessary context. Additionally, I may need to augment anchor transformation to include the crowd's capabilities with cropping or video editing tools.

Finally, while my study has considered one type of error (i.e., conflation error), I cannot draw firm conclusions about other kinds of error patterns that will occur in deployment settings. While I have found no evidence of alternative error patterns, I believe that political, racial, and other biases within the crowd may skew detection results. Prior work have proposed several strategies for mitigating these biases including [104, 185]. Future work should explore applying these strategies in tandem with anchor comparison to control multiple errors.

Despite these limitations, I remain cautiously optimistic. I have tackled one of the key challenges that makes media manipulation challenging to detect (conflation error) and introduced the first approach that can overcome this challenge. While there are likely many more challenges to overcome before my system can be robustly deployed at scale, I believe my fundamental approach is both feasible and important.

3.8 Conclusion

In this chapter, I have taken the first steps toward building systems for social reference processing by contributing a new crowdsourcing approach that transforms a classification problem into a comparison. This allows the crowd to detect text that uses manipulative emotional language to sway users towards positions or actions. My approach, *anchor comparison*, overcomes the challenges that cause automatic and standard crowdsourcing approaches for this problem to perform poorly: the difficulty of gathering comprehensive training data for social references and a tendency for the crowd to conflate emotionally manipulative language (EML) and intrinsically emotional content (IEC). I showed that my anchor comparison approach mitigates conflation errors by transforming the problem into a comparison task where IEC cannot overpower EML. I evaluated my approach by developing a corpus of short text pieces and also showed that my approach affords system builders the ability to tune precision and recall. I argue that my approach could be useful for identifying potential-manipulative content and warning users, helping the public see and understand media manipulation. More generally, my approach is a first step toward solving problems that involve disentangling social references from content.

CHAPTER 4

Justice Panels: Studying Crowdsourced Expertise For Processing Social References

In this chapter, I build on the work from chapter 3 by studying the expertise necessary for understanding social references. I seek to expand the capabilities of anchor comparison by making it possible to identify and recruit a panel of expert crowd workers. I conduct my study in the context of another manipulation task, namely, for understanding the expertise necessary to detect emerging rhetorical narratives that target people-groups based on group identities (e.g., race, gender, disability, etc.). In particular, the capability of detecting these narratives before they are established could aid in crafting effective mitigation efforts. My goal is to make it possible to design a system that overcomes the limitations of my prior system and the limitations other prior systems by recruiting or mimicking the abilities of a panel of social justice experts, which I refer to as *Justice Panels*. In my study, I conducted 27 interviews with experts in social justice where I (a) asked them to highlight and edit three text snippets for racist and prejudiced language, (b) discussed with them what they found, and (c) probed to learn the skills, knowledge, and experiences they used to choose what to highlight and/or edit. My analysis of the interviews has three main findings with important implications for detection systems: 1) that social justice expertise is multifaceted, contextual, and situated; 2) that participant disagreements stem from at least two places including *expertise divergence* and *impact disagreements*; and 3) that *anchor comparison* can be used to maximize each participant's performance. Overall, these findings show system developers need to engage more deeply with panelists' lived experiences in order to build more effective systems.

4.1 Introduction

Rhetoric that others (otherizes) people-groups can lead to the social marginalization of those people-groups. The most impactful forms of such rhetoric are often both highly complex and subtle in nature, avoiding obviously prejudiced words and phrases and basing their main narratives on fragments of truth. For example, rhetoric that impacts immigrants may not target them directly, but instead may create or exaggerate potential costs for American citizens by highlighting cultural differences or by suggesting a grander conspiracy is at play.

Such rhetoric creates a predicament for platforms, which have limited means for consistently detecting and reducing its negative impact. Since this rhetoric often rides the edge between factual description and stereotypical excess, it can evades detection by machines and people alike. Even in cases where it can be detected, it still poses a challenge to platforms, since simply removing the content could be perceived as an overtly political decision. Additionally, othering narratives are extremely difficult to counteract once they are normalized and they can spread and evolve rapidly.

Existing othering detection systems, including both automated and crowd-based approaches, are generally limited to detecting immediately harmful language such as whether specific words and phrases constitute hate speech, toxicity, or microaggressions. While I believe these efforts are critically important, there is still a substantial effort needed to develop technology that can uncover the full complexity of othering argument structures — a requirement for creating impactful mitigation systems. For example, while prior researchers have shown that automated approaches are capable of identifying some cases of othering [8, 28, 33], they are limited to cases where the relevant context is contained within available training data [74]. Likewise, while crowdsourcing-based approaches [12, 39, 194] can process social and cultural meanings in theory, the people the system recruits may lack the necessary expertise to identify those meanings and be capable of understanding the argument the rhetoric alludes to.

A line of crowd system research has made substantial advancements in some of these areas, particularly about improving disagreement recognition and recruitment. Specifically, Kairam and Heer argued that disagreements among crowd responses should be treated as important insights rather than noise [112]. Likewise, Gurari and Grauman's CrowdVerge system expanded on this idea by making early disagreement predictions that make it possible to recruit additional crowd members for high disagreement areas [92]. Finally, Gordon et al.'s Jury Learning system makes it possible to recruit potentially dissenting voices into aggregated crowd decision making by including structural factors (e.g., demographics) in the crowd worker selection process [82]. While these prior studies have advanced disagreement recognition and recruitment systems, it is still not well understood what exactly constitutes disagreement and expertise in a social justice context.

In this chapter, I take steps to expand the capabilities of existing crowdsourcing systems and of othering detection by examining how social justice experts view and find othering rhetoric. I identify the challenges that will be associated with recruiting a panel[1] for uncovering othering and

[1] I use the term *panels* to describe expert groups of crowd workers that are recruited to garner annotation opinions or insights for the purpose of aiding in a decision-making process. Panels are related to, but uniquely set apart from Gordon et al's *juries* [82] where aggregated crowd worker annotations directly control the decision-making outcome.

promoting social justice, and introduce insights that can help overcome these challenges. My goal is to make it possible to design a system that has what I call *Justice Panels*[2] — panels that are either composed of social justice experts or panels that mimic experts capabilities. If I can do this, it would enable sociotechnical systems to rapidly analyze new social media or news content by recruiting people with relevant social justice expertise in realtime, making it possible to discover emerging narrative archetypes that could be used to marginalize people-groups before the archetypes are normalized. Additionally, this approach would make it possible to provide substantial insight into mitigation, such as delivering personalized interventions.

To achieve this goal, I conducted an interpretivist study with 27 people who have substantial expertise in social justice and who are passionate about social justice. My study answers two main research questions: 1) What do participants see as othering rhetoric? and 2) What are the skills, knowledge, and experiences that participants use to find othering rhetoric? To answer these questions, I asked my participants to analyze three text snippets for racist and prejudiced words and phrases while probing them for their thought processes. To analyze my data, I used Clarke's Situational Analysis, an updated version of grounded theory [45]. In my analysis, I observed that my participants found numerous cases of encoded rhetoric being constructed from *semiotic payloads* such as social references, dog whistles, stereotypes, and narratives. I also found that social justice expertise consists of a complex combination of lived experiences, often niche and not tied to a person's identity.

From these findings, I arrive at three design conclusions that have important implications for the design of Justice Panel systems. First, my findings provide evidence that social justice expertise is both *contextual and situated*, which implies that recruitment strategies focusing on simple demographic attributes are insufficient. Secondly, I provide evidence that disagreements between panelists can come from at least two sources, including *expertise divergence* (disagreements due to differences in panelists' expertise) and *impact disagreements* (disagreements about the impact of phrases despite having similar relevant lived experiences). Finally, I show that the *anchor comparison* technique[3] can be used to improve the individual performance of participants.

The rest of this chapter will proceed as follows: I will first describe the relevant background literature (section 4.2) that inspired my understanding of othering rhetoric and the direction of my study, next I will explain the details of my study (section 4.3), then I will cover my main findings (section 4.4), and finally I will discuss the implications of my findings for creating Justice Panels

[2] My intention is to make it possible to create informative panels composed of panelists with relevant social justice expertise. Such panels could provide important insights to inform platform decision-making, such as how to craft mitigation strategies for emerging narrative archetypes.

[3] Anchor comparison is the approach I introduced in 3 that involves breaking down classification tasks into comparison tasks for mitigating the conflation of content with social references. I show below that it can also be used to engage participants with a broader skill set to improve their individual performance.

(section 4.5) and my main conclusions (section 4.6).

4.2 Background

I build upon a foundation of literature on *othering*, a complex social phenomenon studied by a great number of scholars within sociology, psychology, computer science, and related fields. I use the term othering (also known as *otherizing*) to refer to "a set of dynamics, processes, and structures that engender marginality and persistent inequality across any of the full range of human differences based on group identities" [63]. I cannot hope to survey the entire literature, so I focus on *racial* othering[4]. I recognize many other forms of othering exist, such as those focused on gender, sexual orientation, disability status, nationality, and many more.

4.2.1 Racial othering

Racial othering exists in a multitude of forms and permeates almost every aspect of our society, such as within interpersonal relationships [206], news [208], education [15], research [168], public deliberation [148], and clinical settings [202]. For example, To et al. studied how people go online to seek support for personal experiences with racism [206]. They found that finding communities that "get it" is a key challenge for people who are trying to seek support. To et al. also acknowledged the importance of having confidence in the trust and safety of these communities.

While racial othering can take on explicitly hostile forms, it can also take on more subtle forms, such as through microaggression [201], rhetoric [69], or an aversion to particular racial groups [59]. Sue argued that microaggressions have become a substitute for overt racism and are much more common within personal relationships. In their study of microaggressions in psycho-therapist settings, they found that microaggressions are common even among professionals. Ignoring race altogether — known as *color-blindness* — does not preclude acts of racial othering. In fact, it may even prevent progress towards uncovering implicit or systemic forms [25]. While many generally consider contemporary racism to be more subtle [202] than historical forms, scholars have warned that outright cases are still frequent and should not be discounted [147].

4.2.2 Online othering: hate speech, harassment, and toxicity

I focus on the ways in which othering manifests itself within online content, such as content shared within public spaces and fringe communities. Unlike offline communities, which tend to be more

[4] According to many scholars, my definition of racial othering is defined equivalently to racism [53, 187]. However, a few scholars argue that for something to constitute racism, it must also involve a power component. I use this term to encompass all forms of racism and racial prejudice.

secluded, the Internet provides outlets for extremist communities to gather and influence the public [139]. Prior efforts to study the rhetoric within some of these communities found numerous ways in which they otherize marginalized groups, including Meddaugh and Kay's study of the extremist website Stormfront [145] and Baumgarten's follow-up study [18]. These studies showed that racism can manifest not just within explicit language, but through the rhetoric as well. In Meddaugh and Kay's discourse analysis of othering rhetoric on Stormfront, they found that Stormfront appears less virulent than other forums and was thereby more palatable to naïve readers [145]. They demonstrated an example where hate speech is not the main driver for othering, but instead it is "reasonable racism" (within the rhetoric) that can catch a wider audience. They cataloged several ways in which "reasonable racism" manifests, including by making the other seem tyrannical, manipulative, genocidal, inferior, and false martyrs. Baumgarten added to the discussion through their analysis of the language around third parties [18] (groups affiliated with, but not part of the in-group). They found that while third parties can be referred to in an affirmative tone, the overall sentiment is still overwhelmingly negative toward the other.

Racist rhetoric has a large platform impact. For example, Lewis demonstrated in her 2018 study of YouTube that extremist right-wing channels were tightly connected with less extreme channels, providing a social path for viewers to move into more extreme content [132]. Riberio et al. followed up on this study by using a large-scale quantitative analysis [178] to demonstrate that these radicalization paths are in fact active. Likewise, Mathew et al. found that as a result of hate speech going unchecked on Gab, hate speech has become prominent on the network since 2016 [142]. They argued that the engagement that hate speech fosters gives it an advantage over other content due to Gab's algorithms elevating it to the core of its network. Additional studies have drawn similar conclusions about other networks, such as Facebook [114] and Twitter [196].

As a result of the hard work from Lewis and others, platforms have started to identify important concepts that affect users' trust and safety, including hate speech, harassment, and toxicity, among other concepts. Platforms widely use hate speech to refer to language that derides a group of people based on a protected attribute [142], whereas toxicity more vaguely refers to language that encourages people to leave discussions [94]. Online harassment overlaps with these two concepts; however, it refers to a broad spectrum of abusive behaviors enabled by technology and used to target a specific user or users [22]. Platforms' efforts to combat these issues serve a dual purpose: to improve users' trust and safety on the platform, as well as to improve the platforms' public persona for advertisers [75]. While these definitions are useful, I believe that they do not completely cover all cases of othering rhetoric, such as cases that use dog whistles to encode prejudiced beliefs about a people-group and the dangers they supposedly represent to the in-group [198].

In addition to moderation teams, platforms have turned to AI tools to identify this type of content because of the scale of content on their platforms [75, 181]. Specifically, platforms have invested in large content moderation systems that make use of both AI [76, 83] and large teams of people to make decisions about the kind of content they will allow on their platforms. These decisions broadly include deciding which content has misinformation [3], hate speech [142], harassment [22], toxic language [217], and many other decisions that influence users' experience, trust, or safety on platforms, or otherwise represent a brand risk to platforms as a whole [36].

4.2.3 Semiotic payloads

Rhetoric that subtly others people-groups can be communicated through a variety of *semiotic payloads* — devices used to smuggle in additional connotations such as narratives [93], stereotypes [43], tropes [120][5], dog whistles [80], and social references [105]. These semiotic payloads often combine together to form an argument structure — an interlocking chain of premises that lead to a conclusion [62]. Halverson et al. described narratives as a "a coherent system of interrelated stories" [93] that Norambuena and Mitra argued are fundamental to building a perception of the world and have strong rhetorical effect [116, 166]. Likewise, negative stereotypes are often used to describe characteristics about an out-group that could have a negative impact on an in-group (e.g., lazy, violent, dirty) [198].

I argue that a *social reference* — referring to language that invokes connotations by overlapping parts of social and cultural contexts [105] — is a particularly important semiotic payload. Social references can include phrases such as "dirty crime-ridden cities" which invoke a set of racist beliefs for many Americans by referencing a stereotype about minority groups living unclean and immoral lifestyles, and thereby not representing upstanding citizens. Another example includes when the pundit Tucker Carlson used the phrase "living fire alarm" to describe congresswoman Ilhan Omar, which carries the connotation of an emergency due to the threat of an impending fire.

At a linguistic level, social references are terms that make use of a literary trope to create one or more connotations in addition to a more formalized denotative meaning[6]. These overlapping connotations require situational, discourse, and conceptual-cognitive context to decode, concepts Kövecses described in-detail in his catalog of contextual factors that influence metaphorical conceptualization [120]. Additionally, social references can take on the form of a dog-whistle, which refers to messages that are encoded in a way that an in-group can understand but that an out-group interprets with a more moderate meaning [80]. Social references play a key role in contributing to othering rhetoric by conveying both stereotypes and narratives since they can be used to signal social groups, values, or beliefs [43]. For these reasons, I argue that processing social references

[5] Here I refer to *literary tropes*, as in metaphor, synecdoche, irony, etc. [32]. Later in this chapter, I will also use "tropes" to describe recurring motifs, sometimes ones that are cliché.

[6] A useful comparison would be to think of social references like musical overtones, they are connotations that implicitly accompany the canonical meaning.

is key to fully understanding how othering takes place and for creating meaningful interventions.

4.3 Study

In this section, I will describe the steps I took to identify the challenges around creating Justice Panels. Specifically, I designed a study that probes two key aspects of the problem: 1) the components of othering rhetoric and 2) the nature of the expertise needed to uncover othering rhetoric. By studying these two aspects, I was able to identify three main challenges that should be considered in designing systems that will use Justice Panels. I will discuss these challenges in detail in section 4.5. Before that, I will explain my study in detail by discussing the research questions that influenced my design, my dataset and participants, the procedure, and my analytical approach.

4.3.1 Research questions

My study was designed to answer two primary research questions about the challenges in designing a Justice Panels system:

RQ1 What do participants see as othering rhetoric?

RQ2 What are the skills, knowledge, and experiences that participants use to find othering rhetoric?

In answering RQ1 I was able to develop an understanding around how othering rhetoric can manifest using prototypical examples. My exploration revealed a variety of ways that semiotic payloads can build on one another to form rhetorically effective arguments — an important finding to consider in designing a system. Additionally, answering RQ2 helped me understand the aspects of social justice expertise.

4.3.2 Method

To begin my study, I selected three text snippets that contain prototypical examples of racial othering. Two of these text snippets were derived from two articles shared within two popular Facebook groups whose content is largely based around conjuring up fear of Latinx immigrants and Middle Eastern people. My third text snippet subtly targets Black communities and was derived from a popular opinion piece published by a mainstream news outlet. I selected these three snippets because, in addition to racial othering, the underlying facts of all three are largely accurate. My intention was to choose examples where the focus of analysis would be on the language and rhetorical moves rather on the information. From each of these articles, I shortened the content to the length of about 100 words by selecting the most informative paragraphs. My finalized three text snippets range in subtlety and contain a variety of racial othering forms.

Using these three text snippets as probing material, I conducted 27 semi-structured interviews with people who are knowledgeable and passionate about social justice.[7] I recruited participants through snowball sampling, starting with recruitment emails sent to university and corporate email lists. I selected participants based on their extensive volunteer, work, and educational experience within various social justice-related fields, and based on their demonstrated interest in social justice. Additionally, I explicitly recruited some participants based on their identity within the specific marginalized communities that were targeted by the three text snippets[8] (see Table 1 for a breakdown of demographic factors).

4.3.2.1 Procedure

Each interview took place over Zoom[9], where I asked participants to explain their thought processes as they performed two annotation tasks. These tasks consisted of asking participants to 1) *highlight* text for racist and prejudiced language and then 2) to *edit* text to remove racist and prejudiced language. I asked participants to perform each task consecutively on three text snippets (i.e., complete the first task on all text snippets, then complete the second task on all text snippets). Throughout the interview, I probed participants to explain their thought processes, and on completion of the annotation tasks, I asked them to explain why they annotated and what helped them identify specific phrases. Additionally, permitting that time was still available after completing the annotation tasks, I attempted to ask all participants semi-structured interview questions including "what about [their] background was generally helpful in completing the tasks" and "how [they] would explain what to look for". When asked about the factual accuracy of any information in the text snippets, I told them to "assume the underlying facts are accurate".

To do the tasks, I provided participants with a Google Doc containing some instructions, an example of the completed highlight task, and two copies of each text snippet. After explaining the contents of the Google Doc, the general instructions, and the provided example, I asked participants to "highlight the racist and prejudiced language that they see". I instructed them to use Google Doc's built-in highlighting tool to directly highlight the text in each snippet. After completing this task for each snippet, I asked them to return to the first text snippet to "edit the text to remove

[7] One of my participants did not exhibit expertise in their evaluations. I include them in my analysis for completeness, but I limit the use of their data when drawing conclusions from my findings.

[8] I took extra steps to ensure that the identities targeted by the text snippets would be represented within my participant pool, specifically recruiting Latinx people, Middle Eastern people, and Black people. I included these groups since they have a long history of being targeted by racial othering within the United States. However, I believe my findings are applicable to a general range of people-groups.

[9] https://zoom.us/

Participant	Gender	Member of a group targeted by a snippet?	Educational status
P01	M	Snippet 3	Current masters student
P02	F	Snippet 1	Undergrad degree
P03	F	Snippet 2	Undergrad degree
P04	F	NA	Undergrad degree
P05	F	NA	Current undergrad student
P06	F	NA	Current undergrad student
P07	F	NA	Current masters student
P08	F	NA	Undergrad degree
P09	F	NA	Current masters student
P10	F	NA	Current masters student
P11	F	Snippet 2	Current masters student
P12	NB	NA	Current masters student
P13	F	NA	Current masters student
P14	F	NA	Current masters student
P15	M	NA	Current PhD student
P16	F	Snippet 2	Undergrad degree
P17	M	NA	Current PhD student
P18	F	NA	Current masters student
P19	F	NA	Masters degree
P20	F	Snippet 2	Current PhD student
P21	M	Snippet 2	Current masters student
P22	F	Snippet 3	Masters degree
P23	F	Snippet 3	Masters degree
P24	M	Snippet 2	Current masters student
P25	F	Snippet 1	Current masters student
P26	F	Snippet 1	Masters degree
P27	M	Snippet 3	Current PhD student

Table 4.1: Participants' demographic breakdown. Many of my participants have worked or currently work in social work, or have other relevant volunteer experience doing social justice work.

racist and prejudiced language to make it more appropriate for social media". In this second task, participants were able to refer to the highlights they made in the first task and were instructed to directly edit a copy of the text snippet.

I selected my two annotation tasks based on the previous findings from chapter 3 that rewriting can help non-expert crowd workers disentangle social references from content [105]. In that previous chapter, I demonstrated that systems that rely on simple annotation tasks alone (e.g., using only the "highlight" step) can leave non-expert crowd workers vulnerable to conflation error — an error in which workers mix-up content (i.e., core information and facts) for social reference (e.g., emotionally manipulative language). I introduced an approach called *anchor comparison* and demonstrated that it can enable workers to more consistently and reliably do this disentanglement by coordinating workers to construct an "anchor" version of the content that can then be used as a point of comparison against the original. I used the anchor construction part of this approach in the current study to help participants differentiate the racist and prejudiced components of the text snippets from the parts that are informational. My goal with using both tasks was to ensure that participants are able to achieve a consistent depth of analysis across each text snippet.

4.3.2.2 Analysis

I analyzed my data using Clarke's Situational Analysis [45], an updated form of grounded theory [200]. I performed my analysis iteratively, determining themes in the data, recoding my existent data, and collecting new data. Hypotheses about the themes were examined by observing how well they fit additional data, then they were revised on an on-going basis.

My first wave of data collection consisted of participants 1-19, to whom I paid $20 for about an hour to an hour and a half of their time. With consent from my participants, I recorded the Zoom call, transcribed it with Otter.ai[10], and coded it using the software Miro[11]. Within my Miro board, I iteratively searched for patterns between interviews related to my research questions using Miro's post-it notes. I determined themes, discussed the data among the research team, and wrote analytical memos.

After finalizing my analysis of the first wave of data collection, I collected a second wave of data that focused on recruiting participants who identified as a member of the targeted marginalized communities and who had social work experience. To adjust for the additional expertise, I paid these participants an extra $10 ($30 in total). I again recorded, transcribed, coded, and analyzed each interview to find patterns related to my research questions.

The Institutional Review Board at my university approved this study. All data reported here have been anonymized; I have done some light editing of quotes and added emphasis for readability.

[10] https://otter.ai/

[11] https://miro.com/

4.3.2.3 Quantitative measures

To aid in my analysis, I also computed a few quantitative measures to help determine a participant's individual performance relative to other participants and to help identify clusters of participants who annotated text similarly. In particular, after my initial rounds of analysis, I identified features out of the highlighting and editing annotation data (a set non-overlapping phrases that best represent participants' findings). This allowed me to code participant's individual annotation data — which included many overlapping spans of highlighted and edited text — so as to indicate whether or not they caught each phrase after the highlighting and editing steps. In ambiguous cases, where participants highlighted or edited large portions of text, or when they highlighted or edited sub-portions of my selected phrases, I referred to their interview data to determine whether they commented on or made reference to specific phrases. The resulting data consisted of a table representing whether each participant caught a set of specific phrases within the highlighting task or after completing both tasks.

Using this data, I calculated each participant's *recall score*, a measure of the percentage of phrases individual participants caught out of all phrases identified by at least one participant. I used this measure throughout my analysis to tentatively represent participants' individual performance. I found this measure to be reasonable for two reasons: 1) I did not observe any evidence of participants annotating responses "in error" (i.e., annotating text they did not intend to annotate) and 2) after probing participants while they completed my tasks, I determined all annotations to be justified in some capacity. However, I acknowledge that this measure has pitfalls in that it cannot distinguish between types of disagreement between participants — an issue I will discuss in more detail in section 4.5.2.

Additionally, to aid in the analysis of participants' skills, knowledge, and experiences, I identified clusters of participants who annotated a similar set of phrases for each text snippet by using an agglomerative hierarchical clustering algorithm.[12] In the next section, I explain the details of my clusters and I frequently refer to them when answering my second research question.

Now that I have covered the research questions, method, participants, procedure, analysis, and quantitative measures I used to study experts in social justice, I will explain the findings from my study by walking through my data in detail.

[12] Specifically, I created clusters using SkLearn (`https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering`) and following Kairam and Heer [112] by selecting Ward's minimum variance method for linking clusters to form a tree. After referring to interview data, I determined that SkLearn's default suggested split point generated the most sensible participant clusters.

4.4 Results

In this section, I will discuss both what othering rhetoric my participants found and how they identified othering rhetoric. In each of the text snippets, I found that participants identified multiple semiotic payloads that contributed to an overall othering argument structure, including narratives, stereotypes, tropes, and social references. I additionally found that participants made use of a complexity of lived experiences to come to their conclusions. I will begin this section by providing a high-level summary of these findings, and then I will follow with a more in-depth walk-through of the data and my analysis. The full text for each of my snippets can be found at the beginning of each analysis and in appendix A. I summarize my findings below:

Snippet 1: Participants generally arrived at a high level of agreement that the first snippet pushes a narrative that dehumanizes Latino immigrants. To a lesser, but still substantial degree, participants thought the snippet uses stereotypes and tropes to extend the main narrative. The snippet is generally additive, meaning that the phrases participants uncovered combine together to form an increasingly complex argument structure. Likewise, differences in participants' findings can be best explained by the differences in their expertise rather than disagreements about the impact of phrases.

Snippet 2: The second snippet was more encoded and less additive than the first. Most participants agreed that it pushes the narrative that the United States has reached its limit in terms of the number of displaced persons already admitted to the country. However, smaller groups of participants thought that the snippet extends the main narrative by additionally pushing conspiratorial and fear-mongering narratives through the use of encoded social references. I generally found that knowledge of and experiences assisting someone seeking refugee status was most helpful in identifying these narratives.

Snippet 3: My third snippet was the most encoded and least additive of the group. Participants did not agree that the text snippet pushes a specific othering narrative, but almost all participants thought that it pushes at least one stereotype about a prominent Black community. Additionally, I observed that even top performing participants with shared niche lived experiences disagreed with one another.

4.4.1 Text snippet 1

> Six illegal aliens have been charged with sexually assaulting and trafficking a girl under the age of 16 in Wisconsin. All of the six men arrested are believed to be illegal aliens from Guatemala, according to the Green Lake County Sheriff's Office.

> Police said the underage girl called 911 as she was on the run after escaping from the six illegal aliens. After the call, police seized evidence they believes reveals the extent to which the girl was being sex trafficked and sexually assaulted by the illegal aliens. Illegal alien Evis Amabilio Garcia-Rivera is suspected of sex trafficking the girl. Currently, the illegal aliens are being held on bail at the Green Lake County Correctional Facility.

The first snippet is about six individuals charged with sexual assault and trafficking. My participants unanimously agreed that the snippet includes phrases that have an overall dehumanizing effect on the individuals under investigation. The snippet describes attributes about these individuals such as being 'illegal aliens' from Guatemala, the kidnapping and assault crimes they are alleged to have committed, and the current state of the police investigation. In addition to identifying immediate dehumanizing language, participants discussed several ways in which separate phrases add together to create an impactful othering argument structure.

As I have previously discussed, I note that the underlying facts of the story explained in the article are largely accurate — many of the individuals discussed have since pled guilty to charges related to sex trafficking. However, at the time of analysis, many of the facts were still unfolding[13].

I will explain my findings from this text snippet by first walking through a sample of the phrases participants identified, then I will discuss how they combine to create more advanced argument structures, and I will conclude by explaining the lived experiences participants used to identify these structures.

4.4.1.1 RQ1: What do participants see as othering rhetoric?

In total, participants identified 13 unique phrases that contribute to the dehumanizing effect of the text snippet. For the sake of brevity, I select eight phrases that best represent their findings, including 'illegal aliens', 'from Guatemala', 'Evis Ambilio Garcia-Rivera', 'girl under the age of 16 in Wisconsin', 'underage girl', 'is believed to be', 'evidence they believes reveals', and 'is suspected of sex trafficking'. I find that these phrases were *additive* — they combine together to form increasingly more advanced argument structures. I will first walk through the varied explanations participants provided for each term, then I will discuss their additive nature.

Unsurprisingly, all participants thought that the first of these phrases ('illegal aliens') set up a dehumanizing effect; however, their reasoning varied in interesting ways. Some pointed out that usage of 'illegal aliens' dehumanizes undocumented immigrants by exaggerating differences like

[13] Despite my precautions, I acknowledge that the analysis my participants provided might slightly differ from the version they would provide with the latest information about the story. When participants asked, I told them to assume that the facts of the story are accurate.

they are from another planet, while others simply believed that it stereotypes them as dangerous criminals. For example:

> ‘Alien’ to me makes me feel like literally I’m just some weird entity that has come into the country. *It dehumanizes me.* (P02)

And:

> ‘Illegal aliens’ communicates that *the border isn’t safe because people are entering the country in a way I don’t expect.* (P08)

While all participants objected to the phrase, many pointed out that it’s the repeated usage of the term that makes it unquestionably racist (the phrase is repeated six times). A few mentioned that they would be willing to forgive the first or second usage of the phrase had it not been repeated:

> *The first usage isn’t necessarily prejudiced* because if they were undocumented, that that’s a way to describe them. But the choice to describe them that way throughout keeps the focus on that fact. (P10)

From this starting point, the other seven phrases each expand on the initial dehumanization. A majority of participants thought that the first of these (‘from Guatemala’) contributes to racial stereotypes of undocumented immigrants by attributing racial characteristics of Central and South Americans to the crimes committed by the individuals in the snippet:

> *Mentioning Guatemala makes it seem like they’re just kind of profiling that area of the world.* Especially since Central America has many poorer countries and people coming here illegally. (P02)

Others believed that the phrase goes further than just Guatemala and undocumented people. Many believed that the language could be attributed to a larger racial group than simply Guatemalans and the “crime” to more than undocumented entry into the Unites States:

> They’re *typing a stereotype* of people coming into the country, especially from Central America and South America, to being illegal people and involved in this negative thing like sex trafficking. (P10)

Similarly, a smaller portion of the participants thought that the snippet further pushes this narrative by providing the full name of one of the individuals ‘Evis Ambilio Garcia-Rivera’ in a sentence that otherwise communicates no new information:

> It's kind of like saying that people who are named Garcia are perpetrators of this assault. If you're reading the news article and *you don't have any personal experience with people of Hispanic origin, you're going to go "Oh that's a Garcia, so they must be right"*. People internalize this way of thinking. (P27)

In addition to these phrases, about half of my participants objected to the framing around the phrases 'girl under the age of 16 in Wisconsin' and 'underage girl'. These participants thought the phases expand on the dehumanization of the prior two phrases by layering in the old-timey trope "they're coming for our women and children". Participants believed that the text snippet unnecessarily leans into the trope by putting special attention on the woman's age, which they believed exaggerates her perceived fragility and de-centers her from the story to better fit the trope:

> Who's being centered in the story? And, who is this story about? *It's not about the girl, it's about the police fighting crime and it's about saying that these undocumented people are assaulting children.* (P13)

P13 expanded on this sentiment by explaining how media narratives influenced her perception:

> Sometimes in this kind of reporting they'll say 'underage girl' instead of saying child. [...] *If it was an underage boy, they'd be more likely to be referred to as a child*, whereas an underage girl, well, she was 16. (P13)

P12 further explained how the text snippet leans into the narrative:

> There's a different way to talk about the individual that's not just this victim, this *damsel in distress kind of story.* [...] It makes it seem even more polarizing. Like these evil, bad *'illegal aliens' are coming after this underage virginal girl.* (P12)

A few participants thought that the text snippet pushes an expanded version of the trope about Brown and Black men committing sexual violence against white women. These participants pointed out that mentioning that the woman is 'in Wisconsin' highlights the interracial conflict component of the story:

> And you know, they're attacking this innocent girl from Wisconsin. [...] *Wisconsin, I associate with white.* (P03)

Finally, some participants identified language that suggests there may be a lack of evidence confirming the perpetrators' involvement. Specifically, participants mentioned that the phrases 'is believed to be', 'evidence they believes reveals', and 'is suspected of sex trafficking' suggest the snippet is pushing a presumption of guilt based on stereotypes of Latino people:

> I am confused, is it a legit assault? Is it actually confirmed? Have there been charges pressed and everything has gone through the judicial system? *It's basically saying that these men are already charged even without the due process of the legal system because they are 'illegal aliens' and of Guatemalan descent.* (P25)

P16 further explained the significance of these phrases:

> It just sounds like *they're trying to tie typical crimes an undocumented person would commit* even though they don't have confirmation. (P16)

Similarly, a few participants felt that the way the snippet mentions that the individuals 'are being held on bail at the Green Lake County Correctional Facility' is a violation of privacy and further pushes a presumption of guilt:

> I think it is prejudice to give the name of where they are being held. To me, it automatically puts the people at a disadvantage. It's almost like giving their home address, *it's an invasion of privacy*. (P24)

As I have previously discussed, the majority of these phrases were not considered to be racist or prejudiced in their own right or inaccurate statements, but instead, their use is a problem in that they compound to form a dehumanizing narrative: that the people immigrating from primarily Latin American countries have no respect for America's laws and are inclined to commit heinous crimes including sexual violence against vulnerable locals. For this reason, I consider the text snippet to be *additive*.

More specifically, participants found that each of the phrases I have discussed adds a layer to the othering narrative. For example, while the phrase 'from Guatemala' is not inherently racist, about half of participants thought it was inappropriate due to the fact that it is unnecessary information that highlights the race of the immigrants in the text snippet, P03 even clarifying that it "means they're Brown". Likewise, the phrases related to 'girl under the age of 16 in Wisconsin' layer in a sexual violence aspect and focuses the narrative on the immigrants, rather than the girl. These phrases combine with the prior phrase to convert what would otherwise be a more simple "they're coming for our women and children" trope into one of interracial conflict, creating a perceived threat that P08 describes as: "that I need to watch out for men from Guatemala because they might abduct me." Finally, the phases around 'is believed to be' demonstrates the prejudicial component of the narrative, that they are citing the individuals as an example of the narrative without proper evidence. Now that I have described the components that create an othering narrative in the text snippet, I will discuss the skills, knowledge, and experiences participants leveraged to find them.

4.4.1.2 RQ2: What are the skills, knowledge, and experiences that participants use to find othering rhetoric?

In general participants came to substantial agreement in text snippet 1, demonstrating near perfect agreement for the phrase ‘illegal aliens’ and clusters of agreement for phrases related to ‘girl under the age of 16’, ‘from Guatemala’, and ‘suspected of sex trafficking’. Using the clustering approach I explained in section 4.3, I found that participants within similar annotations and expertise were grouped into relatively clean clusters[14] of participants. As I have previously discussed, the first snippet makes use of several tropes and stereotypes about Latinx immigrants to peddle a dehumanizing othering narrative. I will walk through some annotations to show the kinds of expertise my participants drew upon within and between each annotation cluster.

The repeated phrase ‘illegal aliens’ was identified in all of my participants’ annotation clusters. In their explanations, participants reported that they drew upon numerous lived experiences that led them to their conclusions. For example, P02 explained how they felt about the usage of the term “illegal alien” as an immigrant to United States:

> So for me always, as an immigrant in the country, the word ‘alien’ has always been extremely strange to me because it makes me feel like literally *I’m just some weird entity that has come into the country*. (P02)

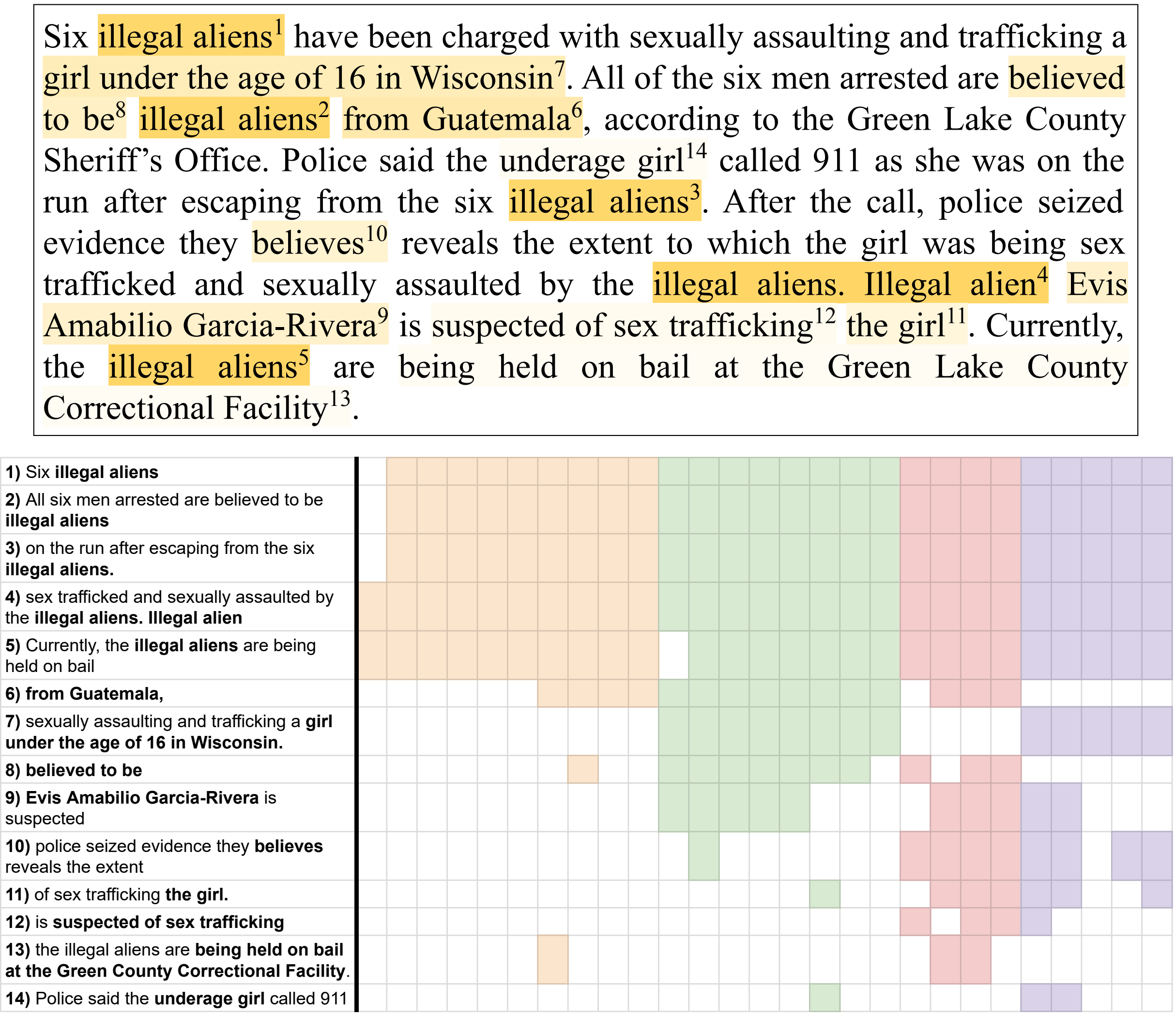
Likewise, P13 recognized that the term was a topic of controversy within the Library of Congress after hearing about it in one of his classes:

> Illegal aliens is not a term that I would use. I think undocumented. There is a whole library of congress controversy recently. *They wanted to take out the term illegal aliens and replace it with I think “undocumented person”* or some version of that, and Congress blocked them from doing that. We talked about it in one of my classes, which is why I was thinking of it. (P13)

Other participants drew upon writing skills they developed through classwork and professional life at large. These participants played with the connotations of the phrase while doing the editing task and recognized that the phrase is simply unnecessary:

> Just by replacing the [phrase ‘illegal aliens’]. *It takes the emphasis off of the men potentially being illegal immigrants and focusing more on the incident that happened and the charges that were brought up [against the men].* I don’t necessarily think that it’s relevant to the charges that are being brought up against them because it’s about sexual assault and sex trafficking. (P10)

[14] By *clusters of participants*, I refer to groups of participants who either highlighted or edited text in a similar way. Please see section 4.3 for more details about how I arrived at specific clusters.



Six illegal aliens[1] have been charged with sexually assaulting and trafficking a girl under the age of 16 in Wisconsin[7]. All of the six men arrested are believed to be[8] illegal aliens[2] from Guatemala[6], according to the Green Lake County Sheriff's Office. Police said the underage girl[14] called 911 as she was on the run after escaping from the six illegal aliens[3]. After the call, police seized evidence they believes[10] reveals the extent to which the girl was being sex trafficked and sexually assaulted by the illegal aliens. Illegal alien[4] Evis Amabilio Garcia-Rivera[9] is suspected of sex trafficking[12] the girl[11]. Currently, the illegal aliens[5] are being held on bail at the Green Lake County Correctional Facility[13].

1) Six **illegal aliens**		🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟪	🟪	🟪	🟪	🟪
2) All six men arrested are believed to be **illegal aliens**		🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟪	🟪	🟪	🟪	🟪
3) on the run after escaping from the six **illegal aliens.**		🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟪	🟪	🟪	🟪	🟪
4) sex trafficked and sexually assaulted by the **illegal aliens. Illegal alien**	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟪	🟪	🟪	🟪	🟪
5) Currently, the **illegal aliens** are being held on bail	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧		🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟪	🟪	🟪	🟪	🟪
6) from Guatemala,							🟧	🟧	🟧	🟧	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟥	🟥	🟥					
7) sexually assaulting and trafficking a **girl under the age of 16 in Wisconsin.**											🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩					🟪	🟪	🟪	🟪	🟪
8) believed to be								🟧			🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟥		🟥	🟥					
9) Evis Amabilio Garcia-Rivera is suspected											🟩	🟩	🟩	🟩	🟩					🟥	🟥	🟥	🟪	🟪			
10) police seized evidence they **believes** reveals the extent												🟩							🟥	🟥	🟥	🟥	🟪	🟪		🟪	🟪
11) of sex trafficking **the girl.**																🟩				🟥	🟥	🟥	🟪	🟪			🟪
12) is **suspected of sex trafficking**																			🟥		🟥	🟥	🟪				
13) the illegal aliens are **being held on bail at the Green County Correctional Facility**.							🟧													🟥	🟥						
14) Police said the **underage girl** called 911																🟩							🟪	🟪			

Figure 4.1: Top shows a heat map indicating the words and phrases participants have highlighted and edited in snippet 1. Phrases are colored by the percentage of participants that highlighted or edited particular phrases, where darker colors indicate a higher percentage of participants that annotated the phrase. Bottom shows a table indicating the words and phrases participants have highlighted or edited in snippet 1. Colors differentiate between groups of participants as determined by hierarchical clustering. The number of participants that identified each phrase is represented by the number of filled-in cells in each row.

In my second and fourth annotation clusters, I can see participants pointed out the phrase 'girl under the age of 16' in addition to the phrase 'illegal aliens'. Again, I observed participants pull from a range of lived experiences related to their career, education, and identity. Specifically, I noticed that participants with lived experience related to women being objectified in the media and broader culture were able to identify the phrase 'girl under that age of 16' and how it is used as a tool for othering by re-centering the story around the immigrants and pushing the "they're coming for our women and children" trope. P12 explained this connection explicitly:

> Being socialized as a female myself, I have been impacted directly by being objectified in that way of being in the assumed victim position of different things, and that *immediately dis-empowers me*. In the first situation where there is a gang of these strong large bodies, it's just a very emotional response. It perpetuates this role of like weak, you know, attaching different things into what it means to be feminine and what it means to identify as a girl or a female. *It's dehumanizing*. They're being stripped of their humanity because she's being treated like property, being treated as a tool to inflict pain or punishment, or drive a point over somebody that's done something wrong. I'm just sensitive to the damaging impact of highlighting gender in this way. (P12)

Other participants leveraged different lived experiences to identify the narrative, such as P13 who showed signs that she used skills developed while working as an editor for a major publisher:

> Who's being centered in the story? And like, who is the story about? *It's not about this girl, it's about the police fighting crime and it's about saying that these undocumented people are assaulting children.* (P13)

Additionally, P17 used his advanced media literacy skills to identify the tone of the snippet:

> Normally, the press, especially if you're talking about white people, would say like "John Smith was charged with sexual assault and human trafficking." *I don't think I've ever seen 'sexual assaulting'*, I don't think I've ever read that before. *It's a minor tone thing.* (P17)

I see similar patterns emerging from my second annotation cluster based around the phrase 'from Guatemala'. Participants in this cluster pulled from numerous lived experiences to identify the phrase, including experiences immigrating to the United States from Latin America:

> By mentioning [Guatemala], in general, *especially for me being from Latin America, feels they're just kind of profiling that area of the world.* Especially Central America, poorer countries and people coming here illegally. (P02)

Likewise, P08 pulled from her experiences as a woman:

> When you read this, *I'm getting the feeling that I need to watch out for men from Guatemala because they might abduct me.* That is a thread of information that you would get out of it and so rewording it in this way doesn't lead me to believe that. (P08)

Finally, I note a group of annotations based around the language 'suspected of sex trafficking'. From this group, a few participants pulled from their identity-based experiences such as:

> I've experienced my fair share of injustice with the justice system and how they do target people of color. And so that's how, whoever wrote this said 'the police identify such and such', *I've experienced that. Being racially profiled and seen as guilty before I prove my innocence.* (P27)

And:

> 'Believes to be', that means that you're working off of perspectives and accusations, you're not basing your evidence on any truths. Maybe also *because I'm a person of color* and reading something like that or also *just living through it* makes me think that most of the time the accusations that lead to detention centers and correction facilities are *based off of someone who believed something, but there was nothing to back it.* (P23)

P25 pulled from her social work experience and her Hispanic background:

> I would definitely say *my Hispanic background and also the social work and social justice lens that was ingrained in us from social work* [was helpful in finding the racist language]. Basically, the repetition of that message that if you are Brown or Black, if you are considered an immigrant or refugee, that somehow you're all put in this pot of "you're bad". And you're either going to do drugs, you're going to be violent, you're going to sex traffic, and sexually assault people. I think that's what kind of made me highlight those things, just that background. (P25)

Participants used a large variety of skills, knowledge, and experiences to identify phrases in text snippet 1. Many of these phrases were obvious, but some were more subtle and required additional expertise to find. In the next text snippet, I will show that participants generally underwent a similar pattern, but expressed more disagreements.

4.4.2 Snippet 2

> A pro-mass immigration organization with links to billionaire George Soros has successfully lobbied six Republican governors to resettle more refugees in their states. The federally mandated refugee resettlement program has brought more than 718,000 refugees to the U.S. since January 2008 – a group larger than the entire state population of Wyoming. Refugee resettlement costs American taxpayers nearly $9 billion every five years, according to the latest research. Over the course of five years, an estimated 16% of all refugees admitted will need housing assistance paid for by taxpayers.

The second snippet discusses an immigration organization that lobbies to expand federal resettlement programs. My participants thought that the snippet describes the organization and their mission in a negative light by making the argument that 'refugee resettlement' comes at a significant cost to American citizens due to the fact that they are 'paid for by taxpayers'. Nearly all participants agreed that the text snippet generally pushes the narrative that the United States has reached its limit in terms of the number of displaced persons already admitted to the country. The text snippet expands on this idea by making use of dog whistles — a kind of social reference — to layer in two additional narratives: first, an antisemitic conspiracy cued by the explicit mention of the Jewish billionaire George Soros, and second, a "replacement of White Americans" one cued implicitly by a comparison with Wyoming. I will again walk through a sample of the phrases participants used to identify these narratives and discuss the various lived experiences participants found to be the most helpful.

4.4.2.1 RQ1: What do participants see as othering rhetoric?

Participants identified 12 unique phrases they deemed to be racist or prejudiced. I present eight of them in more detail including: 'a group larger than the entire state population of Wyoming', 'paid for by taxpayers', 'brought more than 718,000 refugees to the U.S. since January 2008', 'costs American taxpayers nearly $9 billion every five years', 'resettle more refugees', '16% of refugees admitted', 'pro-mass immigration organization', and 'links to billionaire George Soros'. Like in the first text snippet, my participants found that many of these phrases are additive by combining the connotations of some phrases to form new narratives and more advanced arguments.

Most participants mentioned that the snippet makes it seem like immigrants who seek refugee status in the United States are a burden on American citizens and that the country has reached its limit in terms of the number admitted. Most participants identified two specific phrases that push this narrative: the first ('a group larger than the entire state population of Wyoming') highlights the size of the resettlement program and the second ('paid for by taxpayers') highlights the cost of the program. Most participants thought that these characterizations were disingenuous:

> The population of Wyoming is a little small, but since it is a state, I think it brings to mind that this is a ton of people. It's a comparison that *tries to really emphasize how many people that is.* (P04)

Likewise, almost all participants thought that the second characterization that 'an estimated 16% of all refugees admitted will need housing assistance paid for by taxpayers' exaggerates the burden the program has on American citizens. For example, as P14 explained:

> I feel like the reason that they used this was to make the reader think that they will have an effect on their taxes, like *they're responsible for paying for the refugees*. It's almost making the reader automatically think that the *refugees are gonna be a hindrance to them.* It's casting a negative light towards them. (P14)

About half of my participants expanded on this sentiment, mentioning that it is specifically the presentation of these statistics that pushes the main narrative:

> '16% of refugees will need housing assistance' seems pretty low to me. They're kind of like framing in a way like we're going to have to pay for 16% of people who need assistance, but when you're coming into a new place and starting fresh, that seems like a really low number to me. They could have said '84% of people would be able to support themselves'. *They chose to say their statistics in a way that's going to make it seem like a bad situation.* (P10)

This group of participants expressed a similar sentiment toward other statistics in the snippet including how the program has 'brought more than 718,000 refugees to the U.S. since January 2008' and how it 'costs American taxpayers nearly $9 billion every five years'.

> They're trying to make the number '718,000' seem very large by comparing it to Wyoming, but there's like nobody in Wyoming, it's a quiet little place. I would also say *this is sensationalizing the number '$9 billion every five years'*. People are only going to remember the '$9 billion' and not that it's every five years. (P16)

Building on top of the main narrative, the snippet uses encoded social references to layer in conspiratorial and fear-mongering narratives in addition to the main argument. These additional narratives are constructed through the connotations suggested by a few specific phrases, including the way in which the word 'refugees' is used, the phrase 'pro-mass immigration organization', and 'links to billionaire George Soros'. Participants discussed how these phrases combine with prior phrases in the snippet to create new connotations. In general, participants brought up two kinds of supplemental narratives: narratives that make it seem like the United States is being overwhelmed by people admitted through the resettlement program and narratives that allege that there is an on-going conspiracy from political elites to replace American citizens.

About half of my participants expressed concerns over the way in which the word ‘refugees’ is used in the text snippet, suggesting that the snippet’s use of the word has an othering undertone. Specifically, participants thought the phrases ‘resettle more refugees’ when explaining the goals of the resettlement program and ‘16% of refugees admitted’ when explaining characteristics of those admitted unfairly stigmatizes the group:

> I think the word ‘refugee’ is not a bad word per se, but it just *adds a stigma to it that they’re othered*. (P21)

The stigmatized connotations of these phrases has a multiplying effect on the other phrases in the the snippet. For example, the Wyoming comparison develops an entirely new meaning:

> The term ‘refugees’, I wish there was a little bit more of a substantial description of what they’re fleeing from. I think the term ‘refugees’ itself is fine, but when it’s used to refer to a big group of people (or an unidentified group of people) and it doesn’t really tell you what they’re escaping from, it kind of gives off this idea of, like, illegal. Where they’re taking over a region — which is one of the things they’re referring to with the numbers and the ‘larger than the entire state population of Wyoming’. *It’s just this feeling of a big mass that is coming into the country, I feel that’s kind of the vision they’re pushing*. (P04)

Likewise, the snippet creates additional fear-mongering connotations by referring to the immigration organization as a ‘pro-mass immigration organization’. Again, while the language in itself simply appears to be exaggerated, roughly half of my participants pointed out that it combines with other phrases in the snippet to create new connotations:

> It conjures up something sort of fearful, like “these others are coming in”, “they’re replacing us on this widespread scale”, “we’re being displaced”, sort of fear-mongering. Mass immigration suggests some sort of anomaly from historical patterns in the U.S. It’s like *these outsiders are coming and replacing us* and we’re being sort of like polluted or adulterated, and then ultimately disappearing because of these other people who are coming in. I think there’s just these connotations around mass immigration, you’ve just brought in a group of people that’s the entirety of a whole U.S. state. (P15)

These othering and fear-mongering narratives are further expanded through the use of a conspiratorial dog whistle. The snippet describes the immigration organization as having ‘links to billionaire George Soros', a person who is not mentioned again for the rest of the snippet. About half of my participants expressed discomfort with this language, some of them mentioning specifically that the language is a reference to antisemetic conspiracy theories:

> George Soros does fund Open Society Foundation, he does fund work around the world doing great things. But when he's being linked negatively to someone, it does *feel like an antisemetic plot.* And I would say it plays into some very old theories of like, 'who runs the world?' and 'who runs the media?' So that's what that's coded to me. (P13)

Again, these conspiratorial connotations combine with prior connotations to add new meanings to the phrase 'a group larger than the entire state population of Wyoming'. For my participants, in addition to fear-mongering, the conspiratorial angle makes it seem like powerful political elites are involved in a scheme to replace American citizens. Now that I have explained how my participants viewed the many parts of the text snippet that contribute to overall othering rhetoric, I will go over the skills, knowledge, and experiences that participants used to find them.

4.4.2.2 RQ2: What are the skills, knowledge, and experiences that participants use to find othering rhetoric?

In snippet 2, I observe participant clusters based around six cues, including 'paid for by taxpayers', 'a group larger than the state population of Wyoming', 'links to billionaire George Soros', 'pro-mass immigration organization', the use of the term 'refugee', and '$9 billion every five years'. In this snippet, participants leveraged an even greater diversity of lived experiences, resulting in more sparse agreement than from my observations in snippet 1.

Participants in all clusters identified the phrase 'paid for by taxpayers' and nearly all clusters identified 'a group larger than the state population of Wyoming'. Similar to my findings from snippet 1, participants pulled from a large variety of lived experiences to find these phrases, including knowledge of global media narratives, former residences, teaching, and writing skills. For example, I saw P01 demonstrate knowledge of refugee narratives in Germany — perhaps gleaned from his experience as a social worker:

> I think 'a group larger than the entire state population of Wyoming' makes it sound as if *refugees are encroaching, that refugees are a burden on the people who live there*, financially and in other ways. This is the same kind of rhetoric you've heard in Germany and overseas during that whole refugee crisis when a lot of people in Germany are like "no more refugees, we're taking too many in". (P01)

Likewise, I saw P18 pull perspective from living in Milwaukee:

A pro-mass[3] immigration organization with links to billionaire George Soros[5] has successfully lobbied six Republican governors[12] to resettle more refugees[8] in their states. The federally mandated refugee resettlement program[6] has brought more than 718,000 refugees to the U.S.[7] since January 2008 – a group larger than the entire state population of Wyoming.[2] Refugee resettlement costs American taxpayers[10] nearly $9 billion every five years[9], according to the latest research[11]. Over the course of five years, an estimated 16% of all refugees admitted[4] will need housing assistance paid for by taxpayers[1].

1) will need housing assistance **paid for by taxpayers.**		🟧	🟧	🟧	🟧		🟩	🟩		🟥	🟥	🟪	🟪	🟪	🟦	🟦	🟦	🟦	🟦	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
2) since January 2008 **-- a group larger than the entire state population of Wyoming.**	🟧	🟧	🟧	🟧	🟧		🟩			🟥	🟥		🟪	🟪	🟦	🟦	🟦	🟦	🟦	🟨	🟨	🟨	🟨	🟨		🟨	
3) A **pro-mass** immigration organization		🟧	🟧	🟧	🟧	🟩	🟩	🟩									🟦	🟦	🟦		🟨	🟨	🟨	🟨	🟨		
4) an estimated 16% of all **refugees admitted**	🟧	🟧	🟧	🟧	🟧								🟪		🟦	🟦	🟦	🟦	🟦				🟨		🟨		
5) with **links to billionaire George Soros**			🟧			🟩	🟩	🟩		🟥	🟥				🟦			🟦				🟨	🟨	🟨		🟨	🟨
6) The **federally mandated refugee resettlement program**		🟧							🟥	🟥	🟥					🟦	🟦	🟦	🟦								
7) has brought **more than 718,000 refugees to the U.S.**															🟦	🟦	🟦	🟦	🟦							🟨	
8) six Republican governors to **resettle more refugees** in their states.							🟩								🟦	🟦	🟦	🟦	🟦	🟨	🟨	🟨	🟨	🟨			
9) nearly $9 billion every five years, according to													🟪	🟪	🟦	🟦	🟦	🟦	🟦	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
10) Refugee resettlement **costs American taxpayers**										🟥	🟥		🟪		🟦	🟦	🟦	🟦	🟦	🟨	🟨	🟨	🟨	🟨		🟨	🟨
11) the **latest research.**						🟩									🟦		🟦			🟨		🟨	🟨	🟨	🟨		🟨
12) has successfuly **lobbied six Republican governors**																			🟦	🟨		🟨			🟨	🟨	🟨

Figure 4.2: Top shows a heat map indicating the words and phrases participants have highlighted and edited in snippet 2. Phrases are colored by the percentage of participants that highlighted or edited particular phrases, where darker colors indicate a higher percentage of participants that annotated the phrase. Bottom shows a table indicating the words and phrases participants have highlighted or edited in snippet 2. Colors differentiate between groups of participants as determined by hierarchical clustering. The number of participants that identified each phrase is represented by the number of filled-in cells in each row.

> I used to live in Milwaukee and the population was about 600,000. So instead of saying "the population of Milwaukee", they picked this enormous state and *I just happen to know that it has a very, very low population density.* (P18)

Similar to snippet 1, several participants engaged with lived experiences related to general writing skills and experiences with editing during the editing task. For example, P11 made use of her writing knowledge acquired from graduate student instructor duties:

> *I'm a big writer and I also work as a grad student instructor*, so I literally grade hundreds of papers every week. Now that I'm here and reading it with the purpose of rewriting it or editing it, it helps me to see "wait, we actually don't need that second sentence" because this is trying to get us to think one way rather than another. It seems *redundant to the point that it was trying to buy into this narrative of "refugees are bad".* (P11)

In addition to these phrases, participants within the first participant cluster caught the phrases 'pro-mass immigration organization' and 'refugees admitted'. These participants mentioned that they believe these phrases push narratives that induce a fear of being overwhelmed by an influx of immigrants. The participants again pulled from a wide array of experiences such as P14's former experience working in local government:

> I used to work in local government, I worked for a pretty affluent city and the residents were pretty split in what they believe the city should move towards. There are a lot of people who thought that housing should become more dense, more housing should be built, and more people should move in. And then there's certainly the other people. *They identify more with NIMBYISM where it's like "we can't handle more housing."* [The local government] has a news website and if you go to their comment section, they're just very vocal about their views. Sometimes it can get really negative, and there is that whole other ism side to their comments. (P14)

The dog whistle 'links to billionaire George Soros' was the basis of my second participant cluster. In this cluster, P05 mentioned that her Jewish identity was an important factor that helped her identify the phrase:

> What does 'links to billionaire George Soros' mean? Like, is it that he is a donor? It's kind of purposely vague so people can draw their own conclusions, which can oftentimes be more extreme or full of fear than if you just clearly stated what his relationship was to the organization. [...] *I am Jewish and I try to keep myself aware of what people who wish me harm are thinking about.* (P05)

A few participants in other clusters identified the dog whistle as well. A few of these participants also leveraged their identity; however, a few mentioned non-identity based experiences such as learning about George Soros from the news or from memes on Twitter:

> I think I've read about [George Soros] in the news or something. I do remember seeing a picture of George Soros and I think I googled him. *I just remember seeing a number of news things that try to link him to funding conspiracy related things.* (P17)

Two of my participant clusters identified the phrase '$9 billion every five years' and 'costs American taxpayers'. Many of these participants engaged with the numbers provided within the snippet more actively than in other participant clusters. Again, I observe that many experiences were helpful for participants, as P24 explained:

> What was helpful is *my education and social work experience*, just viewing people as truly equal and wanting there to be journalistic integrity. I think my identity plays into it, but I'm not sure how. I think that *it also gives me motivation to see people portrayed unbiased-ly in the media.* (P24)

However, many other experiences were also helpful. For example, P20 used described how her family connections guided her annotations:

> So my husband's from Yemen, he still has family there. Some of his cousin's wanted to come here because of the war, the famine, everything going on, so *we were thinking about applying for refugee status*. There are so many rules, stipulations, and mandates that go along with having refugee status that a lot of people don't want to sign up for that. So the fact that there are 718,000 people signing up for refugee status is huge. Those people all agree to not go back to their homeland, ever. And they agreed to leave all of their friends, family, belongings all behind and come here in hopes of a better life. So they are literally risking it all because they have no choice. (P20)

One of my participants mentioned how they were able to make use of their data literacy skills to make sense of the figures:

> I actually think, the past couple of weeks we've been talking about data literacy in class, so I just have been thinking a lot about how numbers are presented, which is silly. If you're bringing in facts, with numbers especially, I think *people kind of assume more truth because you're saying you have a number* in the sentence, which I know is a little silly. But I think people have a bias towards that as more factual. (P13)

My final participant cluster objected to the use of the term ‘refugees’ in several phrases throughout the text snippet, such as ‘resettle more refugees’ and ‘16% of refugees admitted’. While they did not find the term to be inherently derogatory, they thought that its use in the snippet has a slight othering undertone to it that is used to exaggerate the burden on American taxpayers. These participants arrived at this conclusion through a variety of means, including through the use of skills developed through pre-law and social work experience. One of these participants (P27) demonstrated how skills developed through doing social work helped him think about the snippet from the perspective of a refugee:

> *If I was one of those refugees or one of those people displaced, I wouldn’t want to feel like a burden — it makes you feel like you’re othered.* If you’re already being relocated, you already feel bad for that. So now, we’re disenfranchising them even further by the words we use. We’re telling them: “you’re not an American, you’re a refugee. And now because of you, we have to pay more taxes.” (P27)

However, a couple participants demonstrated some doubt about whether the term should be annotated as racist or prejudiced, such as P25 who decided to change it in the editing task:

> Instead of ‘resettle more refugees’, I would say “assist with asylum”, “assist refugees with asylum”, or “assist individuals with asylum”. *That one’s hard because I see that the word could be helpful, but at the same time, I think ‘refugee’ has been misused so much that people immediately make assumptions about refugees.* (P25)

As I have discussed, participants in text snippet 2 demonstrated a similar pattern of expertise as in text snippet 1. However, I saw some indications of disagreement among participants that were not present in the first text snippet. In the following text snippet, I will again demonstrate how participants continue to follow the general pattern of expertise, but with an example that contains substantial subtlety.

4.4.3 Snippet 3

> Caldwell’s story is remarkable and inspiring. One of nine children raised in poverty on the South Side of Chicago, he had to face true adversity. His mother was addicted to crack. His community was poor, rundown and violent. Caldwell recalled to me how he used to think Republicans belonged to the party of racists who didn’t care about minorities or the poor. But then one day, an elderly Black man on the South Side challenged him, explaining that at one point most Black Americans were Republicans and conservatives. So, Caldwell did a little research and realized the man wasn’t lying.

> So began his transformation to become a Black conservative. Interestingly, Caldwell explained that many Black Americans are actually conservative and believe in family values – they just don't vote for Republicans.

So far, I have seen snippets that carry their racism overtly and obviously, with flourishes of more covert language. However, this is not always the case. The final snippet in my dataset has much more encoded language, and it rides the edge between a factual description and stereotypical excess. The snippet had the least amount of agreement from my participants about which specific words and phrases are racist or prejudiced. Some participants did not even view the argument structure as inherently racist.

The text snippet describes the experiences of a man named Caldwell who supposedly grew up in a poor Black community. The snippet alludes to some of the conditions Caldwell endured growing up including drug abuse and violence. It transitions to telling a political coming-of-age story, which was fostered by a conversation with an elderly community member who pointed him toward conservatism. It concludes by making the argument that many Black Americans hold conservative values, but have yet to start voting for Republicans.

The othering in the snippet is created through a complex rhetorical structure: there isn't a single narrative that participants all agreed upon, but instead, the othering manifests through a fragmented mashup of many narratives. A sample of these narratives include: that Black communities are dilapidated due to poor choices, that members of these communities can overcome their hardship if they so choose by pulling themselves up by the bootstraps, and that doing so requires a transformation into conservatism. These narratives are formed through the use of multiple stereotypes, dog whistles, and tropes throughout the text snippet. While some of these narratives were debated by participants as to their racism or their truth, nearly all of my participants identified some aspects of the snippet they thought to be racist or prejudiced. Some participants even felt that the racism is the most impactful in this snippet when compared to the prior snippets. As before, I will walk through some of the phrases participants used to identify these narratives, then I will discuss the various lived experiences participants pulled from.

4.4.3.1 RQ1: What do participants see as othering rhetoric?

Overall, participants called attention to 15 unique phrases they believed to contribute to the racism or prejudice within the text snippet. I will cover a sample of these phrases, including: 'addicted to crack', 'poor, rundown, and violent', 'South Side of Chicago', 'Republicans belonged to the party of racists who didn't care about the minorities or the poor.', 'transformation to become a Black conservative', 'Caldwell's story is remarkable and inspiring', 'elderly Black man', and 'family values'. The implications of these phrases were complex. I will discuss the diverging participant perspectives for each.

Most of my participants said that they believe the snippet pushes negative stereotypes about Black communities. In particular, participants identified two places in the snippet: the first one where it describes Caldwell's mother as being 'addicted to crack' and a second instance where it describes Caldwell's community as being 'poor, rundown, and violent'. I note that while participants identified several stereotypes present, participants varied greatly in what role they believed the stereotypes play in the rhetoric of the snippet. For example, the description of Caldwell's mother's addiction was viewed by some to be just a derogatory way of referring to his mother:

> I feel like it's a very insensitive way of wording it. They could have just said 'a substance problem'. *Calling it 'crack' is almost crass.* (P14)

Other participants felt that the language is phrased in such a way that it serves a greater purpose than to just be derogatory; it pushes a negative stereotype about Black communities:

> They're not saying that he's African American, but they let you know that by saying things like: 'one of nine children raised in the South Side of Chicago', which is code for "he's Black", and 'he had to face true adversity', and 'his mother was addicted to crack'. *That's what a lot of White America thinks of Black people, this is the image.* (P03)

Some of my participants felt even more strongly about the usage of 'crack'. They specifically brought up that the term carries negative connotations from the war on drugs:

> I think 'crack' *is a very loaded topic and term that has been very overused and refers to a real epidemic.* I think you could have said "his mother was an addict" without saying 'addicted to crack' and that would have meant something more neutral. It has racial connotations even though historically there was a crack epidemic. I think culturally, it's come to be like a very loaded topic. (P13)

Likewise, most of my participants thought that the language 'poor, rundown, and violent' used to describe the South Side of Chicago boils down to a stereotypical depiction:

> We don't need to say that 'his community was poor, rundown, and violent'. That is, sort of, *the stereotype we think of when we think of the South Side of Chicago*, so there's no point in re-emphasizing that. It pushes the narrative that he grew up in this really bad, poor, Black community. (P11)

P17 further explained how the words 'rundown' and 'violent' contribute to the negative tone:

> I would say that a lot of people come from poor communities, but to say that the community was not only poor, but it was also 'rundown' and 'violent'. It's like adding negative words to strengthen the point. And also just to say that a place is violent, rather than to say a place experiences violence. *The place itself is not intrinsically violent, if there is violence in the place. It's a subtle tone thing.* (P17)

About half of my participants felt that mentioning the 'South Side of Chicago' has particular significance when using such language to describe the community:

> I feel like *the South Side of Chicago is known as this rundown area, a lot of Blacks live there.* And it's just like, not well off. I think I hear about a lot of shootings and things that happen there and I think they're just trying to paint this picture, *really emphasizing that stereotype of where this person came from.* There's a lot of things that shape a person and they're choosing to use these few things. (P10)

The use of these two stereotypes (drug addiction and dilapidation) to describe the 'South Side of Chicago' combine to form more intense racial connotations. In particular, some participants believed that the combination makes it seem like Black communities are *inherently violent*:

> I think typically when I see a combination of words like 'poor' and 'violent', and 'rundown', that's always a red flag to me because it's usually a *very reductive perception of what is going on in a neighborhood or within a class of people*. It's this idea that criminals only come out of poor neighborhoods. It's indirectly racist, like *this race of people is this certain way.* (P08)

P05 expanded on this by describing how the phrases form a narrative about Black communities:

> I think using the 'South Side of Chicago', I've seen that brought up when fear wants to be inspired. Like, oh it's so violent and poor and also largely non-White. And I think some of how racism works is focusing on other things that are wrong with this situation *without saying explicitly that it's this community full of Black people. Even though it could be all true, I think it's more used as a tool.* Rather than just listing the facts about his life, I think it's used as a tool in the narrative of the whole paragraph. (P05)

In addition to raising the stakes of the stereotypes, the phrases serve another combinatory purpose: to form the narrative that Caldwell pulled himself up by his bootstraps to overcome his hardship. P16 explained how they arrived at this conclusion:

> I feel like all of these success stories I hear — like I've read in college, on the news, or on television or something — *gives me a "pull themselves up by the bootstraps"* feeling. [...] I guess I don't love how the Black narratives that are being told in mass media. It all kind of sounds the same. (P16)

Additionally, participants discussed three more components to the story that reinforce the ideas of these narratives. The first is a stereotype about politics and race that pushes a reductive view of what constitutes racism, the second is a reinforcing narrative about turning from one's evil ways and coming to enlightenment, and the third is a dog whistle that suggests the Republican party has the moral high ground. I note that many of my participants thought that these components were not outright racist on their own. However, they combine together to form a complex and powerful argument structure.

The first of these components is a negative stereotype about the Republican party that Caldwell allegedly once believed: 'Republicans belonged to the party of racists who didn't care about the minorities or the poor.' At first glance, it may appear that including this stereotype is a concession, as a few of my participants expressed:

> That seems odd but on the other hand it's not saying that's what Republicans are, it's just saying he recalled that he used to think that. *This is just someone's perspective in life.* (P02)

However, about half of my participants noted that the stereotype is important for the argument of the snippet. In particular, the stereotype boils racism down to simply differences in partisan politics, and then writes it off entirely later. P12 explains this:

> This is a prejudiced thing about the Republican party that I've seen expressed in different articles and media, trying to align racism strictly with the Republican party, maybe to influence where people of color are voting or not. It's *attaching the idea of racism to a specific group of people and how they identify.* So if I were to identify as Republican or a member of the Republican party, then this statement would be saying that I am racist and don't care about minorities or the poor. (P12)

In the second component, the snippet uses a Constantine trope[15] that Caldwell went through a 'transformation to become a Black conservative'. The trope piggy-backs off of the stereotypes about Black communities to layer in the idea that members of these communities can overcome their adversity if they undergo a transformation to conservative ideology. This ideology would

[15] I refer to the Roman emperor Constantine who is claimed to have miraculously seen a cross in the sky and, as a result, converted to Christianity.

allow Black community members to see that if they would simply work harder, they would be able to overcome their adversities. The trope is established in two places in the snippet: phrases about how ‘Caldwell’s story is remarkable and inspiring’ and a trope about being challenged by a wise ‘elderly Black man’.

The first of these phrases, ‘Caldwell’s story is remarkable and inspiring’, sets up the trope for later in the snippet. The snippet completes the trope by describing how Caldwell ‘began his transformation to become a Black conservative’, which many participants interpreted as an implication that transforming to a ‘Black conservative’ is what is ‘remarkable and inspiring’:

> It’s not saying, he became the CEO or he did good for his community, political stance aside, it’s *“he went on to become a Black conservative” which allows people to tie ‘remarkable and inspiring’ to [being a] conservative.* So conservative is what you want to do, what you want to achieve, because that is what’s inspiring and you want to inspire people. (P20)

Some other participants were cued into the trope through another common element: a wise elderly man who helps a person see the errors in their ways depicted through ‘an elderly Black man on the South Side challenged [Caldwell]’. These participants thought the phrase established credibility for Caldwell’s ‘transformation’:

> From what I understand of Black culture is that there are a lot of similarities in Arab culture and Black culture in that you always respect the elderly. *You always listen to the elderly and take what they have to say as words of wisdom.* It wasn’t just “a Black man challenged him”, or “a Black kid challenged him”, it was elderly Black man. (P20)

Other participants focused their criticism on the in-group identity of the man, mentioning that the snippet is using his identity to push an agenda without actually letting him speak. P16 explains how they came to this conclusion after viewing Candice Owens videos about George Floyd (many of my interviews took place near the 2020 election):

> What research did [Caldwell] do? What did this man say to [Caldwell]? I’m just expected to believe that’s true. I had a hard time seeing all the Candice Owens videos through this election cycle with everything happening with George Floyd around that time and all the White people posting it. Just because one Black person doesn’t believe racism is real or that they’re not oppressed or haven’t been oppressed historically doesn’t make them a monolith. So this kind of story is like, you’re entitled to feel the way that you feel, but that *doesn’t mean you speak for all Black people and that doesn’t actually mean that your experiences are the truth.* (P16)

Finally, the snippet also contains the dog whistle ‘family values’ which is used to push the idea that conservatives hold more righteous values than liberals do. About half of my participants identified the ‘family values’ dog whistle, although interpretations varied by participant. Some of participants found that the dog whistle is simply making some specific values seem more likable than they may be on their own:

> Family values was something that was kind of *codified for hetero-normative*, like to man, wife, nuclear family. (P04)

Other participants thought the dog whistle is intended to target other communities by making it seem like conservatives are the only civilized ones:

> So this is just the opposite of saying that Black people believe in family values, as if they don’t normally. I think *it’s the kind of family values that Republicans have appropriated.* The myth is that they care about families and ideas like being honest and doing hard work and going to church, don’t have abortions and are not gay. (P19)

As I have explained, participants found multiple stereotypes, dog whistles, and tropes within snippet 3 that they believed create othering rhetoric. I will now discuss the skills, knowledge, and experiences they used to find this phrases.

4.4.3.2 RQ2: What are the skills, knowledge, and experiences that participants use to find othering rhetoric?

In snippet 3 I observe the largest annotation variance of the snippets. There are two sources of this variance that influenced three participant clusters. I will again present my data by walking through my interview findings for meaningful differences among participant clusters.

Like in prior snippets, I see a few phrases that were caught by nearly all participants, including ‘his mother was addicted to crack’ and ‘his community was poor, rundown, and violent’. Again, participants pulled from their lived experiences to come to their conclusions, like P27 who found the phrase ‘addicted to crack’ degrading based on how he would feel if someone were to refer to his own mother in the same way:

> I have experienced this, but I hate when people write, excuse my label, bullshit like this. Because it’s like, while I get what you’re trying to say, there’s other ways of saying it. It’s so degrading. *I know what my mother did, but I don’t want nobody talking about her like that.* (P27)

Caldwell's story is remarkable and inspiring.[15] One of nine children raised in poverty[11] on the South Side of Chicago[10], he had to face true adversity[12]. His mother was addicted to crack[2]. His community was poor, rundown and violent[1]. Caldwell recalled to me how he used to think Republicans belonged to the party of racists who didn't care about minorities or the poor[5]. But then one day, an elderly Black[7] man on the South Side challenged him, explaining that at one point most Black Americans[13] were Republicans and conservatives. So, Caldwell did a little research and realized the man wasn't lying[14]. So began his transformation to become a Black conservative[6]. Interestingly, Caldwell explained that many Black Americans[8] are actually conservative[4] and believe in family values[3] – they just don't vote for Republicans[9].

1) His community was poor, rundown and violent.	🟧	🟧	🟧	🟧	🟧	🟧					🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥		🟪	🟪	🟪	🟪	🟪
2) His mother was addicted to crack.		🟧	🟧		🟧	🟧		🟩			🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟪	🟪	🟪			
3) are actually conservative and believe in **family values** -- they just don't vote for Republicans.				🟧	🟧	🟧			🟥	🟥	🟥		🟥	🟥	🟥	🟥	🟥	🟥		🟥				🟪		
4) many Black Americans are **actually conservative**					🟧				🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥						
5) Caldwell recalled to me how he used to think **Republicans belonged to the party of racists who didn't care about minorities or the poor.**							🟩	🟩	🟥	🟥				🟥	🟥	🟥	🟥	🟥	🟥	🟥					🟪	🟪
6) So began his transformation to become a **Black conservative.**									🟥	🟥	🟥			🟥		🟥	🟥	🟥	🟥	🟥					🟪	
7) But then one day, an elderly **Black** man on the South Side challened him									🟥		🟥					🟥	🟥	🟥	🟥	🟥						
8) Interestingly, Caldwell explained that many **Black Americans**									🟥	🟥			🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥			🟪	🟪		
9) — they just don't vote for Republicans.										🟥		🟥	🟥	🟥	🟥	🟥	🟥	🟥					🟪	🟪	🟪	🟪
10) One of nine chilrdren raised in poverty **on the South Side of Chicago,**											🟥	🟥	🟥		🟥				🟥	🟥	🟪	🟪	🟪	🟪	🟪	
11) One of nine children **raised in poverty**											🟥		🟥					🟥	🟥		🟪	🟪			🟪	
12) he **had to face true adversity.**											🟥	🟥		🟥	🟥		🟥	🟥	🟥	🟥		🟪		🟪		
13) explaining that at one point most **Black Americans** were Republicans and conservatives.																🟥	🟥	🟥	🟥	🟥		🟪	🟪	🟪	🟪	
14) did a little research and realized the man **wasn't lying**.											🟥	🟥				🟥	🟥	🟥					🟪	🟪	🟪	
15) Caldwell's story is remarkable and inspiring.	🟧										🟥	🟥				🟥				🟥						

Figure 4.3: Top shows a heat map indicating the words and phrases participants have highlighted and edited in snippet 3. Phrases are colored by the percentage of participants that highlighted or edited particular phrases, where darker colors indicate a higher percentage of participants that annotated the phrase. Bottom shows a table indicating the words and phrases participants have highlighted or edited in snippet 3. Colors differentiate between groups of participants as determined by hierarchical clustering. The number of participants that identified each phrase is represented by the number of filled-in cells in each row.

However, unlike in prior snippets, I observed indications that some participants disagreed with one another about the impact of this phrase. Specifically, P26 came to a different conclusion about the phrase despite having similar social work experiences within Black communities and a personal experience of having a parent suffer from substance abuse. While both participants (P27 and P26) acknowledged the demeaning connotations of the phrase, P26 expressed how she believes Caldwell has the right to describe his mother how he chooses:

> I try to refrain from editing someone's experience. My dad's an alcoholic, I wouldn't be like "oh my dad had substance use disorder when growing up", I would be like "my dad's an alcoholic." You know, this guy's okay with being like "my mom was addicted to crack", not "my mom had substance use disorder". I work in [an intercity] now and a lot of [my clients] would say, "yeah my mom was addicted to crack". So I understand where that's coming from, like the connotations of it and the stigma of it. But also, *if that's how Caldwell is saying it, then I don't really feel like I have to change it.* (P26)

Participants also indicated that the phrase 'South Side of Chicago' is racist; experience was not the only factor that influenced participants' conclusions. For example, P22 expressed that she thought the language plays into a typical narrative she has heard throughout her life that overlooks Black Americans:

> I grew up in [a college town], so in that aspect I did live in a very nice home. I have family that live on the South Side of Chicago, just like I have family that live in [the suburbs] and I think maybe [my demeanor] just changed [while reading this one] because I feel like *this is just the typical story that I hear all the time and it infuriates me.* Honestly, with [my local college] though, they are a very White superior type of school, and I felt it when George Floyd passed away. So I think anything like this, especially when it comes to Black Americans, I think I get more upset by that. (P22)

However, while many participants were able to explain the stereotypes around the South Side of Chicago, they disagreed that the snippet itself is racist. In contrast to the explanation from P22, some participants thought of the phrase as more of a statement of fact than as racist or prejudiced on its own. These participants described several ways in which they learned about the South Side, including from personal job experiences and reading. For example:

> I just know that the South Side of Chicago is a predominantly Black neighborhood that is more known to be poor. I did an interview at a school there once, and I guess the school, *they sort of prepared us beforehand and said "just be really careful when you're in our neighborhood".* So it's something that I kind of knew beforehand. (P14)

One cluster of participants was uniquely set apart from the others in that participants identified the phrases ‘family values’ and ‘actually conservative’. Many participants in this group felt that the phrases are used as a dog whistle to signal conservative talking points while remaining under the guise of uncontested American values. P17 explains how his experiences living in a conservative community led him to this conclusion:

> I grew up in a rural farming community in Michigan. I grew up around a bunch of Republicans and ‘family values’ is just a weird thing. In the last couple of years, probably around the 2016 election, there were “defend religious freedom” signs in people’s yards. It’s like, religious freedom is free, that’s not even on this political discussion table. It’s an innate right. *What they’re really saying is “prevent abortions because Catholics think that they shouldn’t happen”*, you know? In the same way it’s co-opting, religious freedom and family values are these GOP talking points to sugarcoat this weird idea where we want people to have the freedom to say what they want and buy guns, but we want to restrict people’s freedom for gynecological reproductive decisions. (P17)

Unlike in previous snippets, some of my participants mentioned how they didn’t object to specific language elements of the snippet. However, they did object to the overall narrative around “overcoming adversity”. These participants pulled from experiences related to their own relationship with adversity and their education. For example, P22 brings up her experience as a Black woman growing up in a White household:

> I think because we knew he was a Black man, it was just easy to start talking about how he faced adversity. And I don’t want to diminish that, because I’m sure he did, but also, I’m sure he’s so much more than the adversity he faced. And you know, I don’t know his story, I’m sure the adversity helped him become who he is. Let’s talk about his accomplishments and who he is as a person, not the barriers. [...] I’ve also been told “how lucky you are to grow up in a White household.” So I’m like “no, I faced a lot of it”. *Yeah I was lucky, and I also faced adversity growing up in an all White household, but that makes me me and I can’t dwell on that.* And you know, if anything, that’s my fight. (P22)

Many participants specifically described how challenging determining racist and prejudiced language was for them in this snippet as compared to the others. Participants pointed to many considerations that made it challenging for them to complete the task, including the structure of the racism in the snippet and the relationship of the author. For example, P02 described the racism as manifesting itself through clichés, which made it challenging for her to identify specific language:

> This one is odd because it's kind of recounting someone's life. I mean "raised in poverty on the South Side of Chicago", every time you hear about the South Side of Chicago people link that to poverty. But that's, I guess, the reality and story of many people who come from that area, so I'm not sure if that's pointing out something that shouldn't be said. Or "his mother was addicted to crack", I don't know how much that adds to the value of the text. You almost make the connection "oh you're poor and you live in a rundown and violent community? Oh so there's someone addicted to crack?" *I don't see anything outright racist about it, but it seems to be hitting all of the kind of cliché points that help people make these assumptions about other people.* (P02)

Likewise, P24 found it difficult to identify racist and prejudiced language because of how central much of the language is to the core story in the snippet. He did not believe the same was true in the prior snippets:

> *I'm having a tough time changing some of this because I don't want to take away from the core part of the story.* It's different from the last [snippet] because with the last [snippet], they had a lot of extra information in there whereas this is core to the story. (P24)

Specifically, participants brought attention to Caldwell's role in the story and how challenging it was to pin down whether his experience is being co-opted for another. For example, P05 describes her challenges determining the ownership of his narrative:

> It's also like, how much of a part is Caldwell taking a role in the writing of this story? I don't know if this is just someone who was able to find these facts about him or actually interviewed him. Because this could be his own narrative, which is also I think, a part of the reason why *I have an aversion to changing anything, because I also don't want to be someone who tells groups how they should tell their own story.* (P05)

Likewise, P25 expressed doubts that snippet accurately depicts Caldwell's experience:

> It takes his experience to mean something else. [...] *I just don't like the co-opting of his story to make it seem like he's finally seeing the light at the end of this tunnel of his adversity and that light is just becoming a conservative.* Not like, trying to earn money to get out of the situation or helping his community in some way. (P25)

I have just walked through three text snippets that represent prototypical examples of othering rhetoric where I discussed the various ways in which individual phrases combine to create othering argument structures. Additionally, I walked through the variety of skills, knowledge, and experiences that were necessary to find these phrases. In the next part, I explain three follow-up analyses I conducted to summarize my overall findings and to further investigate my research questions.

4.4.4 General findings

I conducted three follow-up analyses intended to explore what influenced participant responses. In the first of these follow-up analyses, I took a closer look at what was helpful for participants when completing my tasks for each snippet. Secondly, I conducted a follow-up analysis intended to uncover why participants disagreed with one another. Finally, I examined how the two tasks I asked participants to complete (highlighting and editing) impacted their performance.

4.4.4.1 What made participants "top performers"?

While all of my participants are experts in social justice, some appeared to have even more expertise than others. The top-performing participants in my study — participants that identified the most unique racist and prejudiced phrases and scored in the top 20% of recall scores[16] — made use of a huge variety of skills, knowledge, and experiences when completing my tasks. Very generally, these participants demonstrated an exceptional breadth of *relevant lived experiences* (often stemming from their volunteer service, work, or personal life) and *media literacy* (including knowledge of media narratives, social justice issues, and journalism skills). I witnessed that participants with both of these components were usually the top performers, while average performers (middle 20-80% of recall scores) often had only one component, and bottom performers (bottom 20% of recall scores) generally had neither. I will explain this finding in more detail by examining the top performers in each text snippet.

In text snippet 1, the top-performing participants caught both obvious and non-obvious phrases including phrases related to 'from Guatemala', 'girl under the age of 16', and 'suspected of sex trafficking'. While these participants brought up many kinds of lived experiences, nearly all specifically mentioned experiences related to baseless accusations and sexual objectification — some having been targeted by and others having kin who were targeted. They also demonstrated significant media literacy garnered through social work experience, education, or volunteer service that enabled them to connect the text snippet's main narrative with media narratives that objectify women to push stereotypical representations of minority groups. Average-performing participants generally expressed less relevant lived experiences and demonstrated fewer examples of their media literacy skills. In all cases, I see participants often arrive at the same conclusions despite coming

[16] As explained in section 4.3, by recall scores I refer to the percentage of phrases an individual participant caught out of all the phrases any participant caught.

from different directions, such as P12 who used their gender experiences to find ‘girl under the age of 16’ while P13 and P17 used their knowledge about how media typically reports cases of sexual assault to find the phrase.

The top performers in text snippet 2 similarly used many relevant lived experiences and media literacy skills to identify most of the statistics in the snippet, ‘a group larger than the entire state population of Wyoming’, and phrases related to the use of ‘refugees’. However, almost all of these participants specifically made use of immigration-related experiences such as assisting relatives, friends, or clients who have or are going through the process — only one had gone through the process themselves. Additionally, I observe a few of these participants make use of their media knowledge to identify the phrase ‘links to billionaire George Soros’ and others use their education and social work experience to find ‘$9 billion every five years’. Similar to snippet 1, average performing participants tended to express fewer examples of relevant lived experiences and media literacy skills. However, even among these participants, immigration-related experiences were a useful tool for finding phrases.

Text snippet 3’s top-performing participants made use of the widest range of lived experiences and media literacy skills to identify phrases. For example, P17 made use of their experiences living in a rural farming community to find ‘family values’, P22 made use of their experiences as a biracial person to identify ‘had to face true adversity’, and P27 made use of their experiences dealing with family overcoming drug abuse to identify ‘addicted to crack’. Likewise, they again demonstrated substantial media literacy skills to connect the phrases to stereotypical media narratives, such as the narratives around Black folks “overcoming adversity” and the use of ‘family values’ to represent conservative policy.

Across all of the text snippets, I find that a few participants generally scored well on all text snippets, but most varied by text snippet. Specifically, three participants (P22, P25, and P27) were the top performers for all text snippets while the other two slots changed depending on the snippet. I are able to identify some of the skills, knowledge, and experiences top-performing participants contributed, but many of them were niche and cannot be easily categorized. Importantly, while a participant’s identity (i.e., racial and gender) often helped them identify phrases, there were many other relevant skills, knowledge, and experiences that helped them, such as work, education, volunteer, and other personal experiences. For example, while P20 had not gone through the U.S. immigration system herself, her experience aiding family members through the process of obtaining refugee status helped her identify numerous racist and prejudiced phrases in snippet 2. Additionally, I repeatedly found that skills related to a person’s media literacy such as their knowledge of media narratives and journalism skills were important for finding phrases — especially when participants lacked relevant lived experiences. Some participants used these skills extensively, such as P17 who was able to connect many of the phrases to popular media narratives.

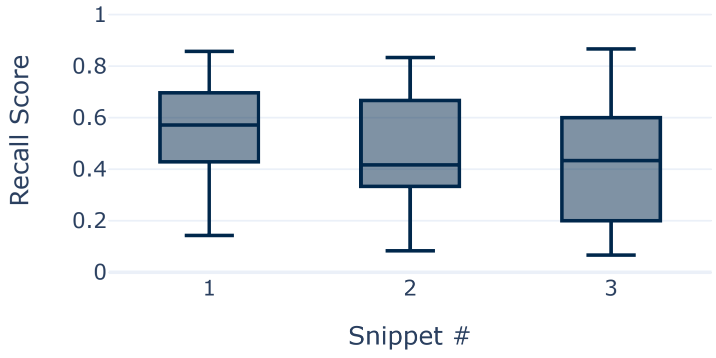


Figure 4.4: The distribution of recall scores across all three snippets. Average scores are highest in snippet 1, and decrease in snippet 2 and 3. Standard deviations follow the opposite pattern: they are the lowest in snippet 1, higher in snippet 2, and the highest in snippet 3.

4.4.4.2 What made participants disagree with one another?

Based on my analysis of the participants' responses, I believe my participants disagreed with one another in many instances throughout each of my text snippets. Two types of disagreement were observed: *expertise divergence* and *impact disagreements*. Expertise divergence is where participants exhibited different levels of expertise and thereby were able to identify different phrases. In contrast, impact disagreements consist of situations where participants have a similar level of expertise in terms of relevant lived experiences, but interpret the impact of phrases differently.

I observed many cases of expertise divergence in text snippet 1, where I found four distinct groups of participants that each additively identified phrases. I uncovered substantial evidence that these differences in annotations were primarily due to differences in participants' skills, knowledge, and experiences. For example, some of my participants mentioned that their experiences related to sexual objectification helped them find phrases related to 'girl under the age of 16', while a few others used experiences related to baseless accusations to find phrases related to 'suspected of sex trafficking'. A few other participants arrived at similar conclusions about these phrases but used their knowledge about how women and minority groups are often portrayed in the media.

In text snippet 3, I observed indications of impact disagreement where participants disagreed about the impact of specific phrases despite having similar relevant lived experiences. I saw this particularly with the example of whether Caldwell has the right to describe his mother as 'addicted to crack' where I found two participants — both with marginalized identities, social work experience working in Black communities, and personal experiences with a parent going through substance abuse — arrive at different conclusions about Caldwell's intentions. Similarly, with phrases like 'on the South Side of Chicago', most participants agreed that it sets up a stereotypical description of Black communities (again playing into the "overcoming adversity" narrative), but disagreed about whether including the fact itself is racist or prejudiced.

These disagreements can also be observed quantitatively by examining the distribution of participants' recall scores — the percentage of phrases an individual participant caught out of all the phrases any participant caught. As can be seen in Figure 4.4, the total level of disagreement increases from snippet 1 to snippet 3. Participants caught the highest portion of phrases the most consistently in snippet 1 (mean=0.54, std=0.17), a smaller portion of phrases less consistently in snippet 2 (mean=0.48, std=0.23), and the lowest portion of phrases the least consistently in snippet 3 (mean=0.44, std=0.26). Additionally, using the *Silhouette Coefficient*[17], I observe that participants were able to be grouped into clusters the most neatly in snippet 1 (0.31), the next most neatly in snippet 2 (0.20), and the least neatly in snippet 3 (0.18).

4.4.4.3 How did my annotation tasks impact participants' thought processes and performance?

I found evidence that the type of tasks participants completed can impact their individual performance. Specifically, nearly all participants identified phrases in the rewriting task that they did not identify in the highlight task (see Figure 4.5). While participants increased their recall scores by 12% on average after completing both tasks as compared to completing the highlight task alone, some participants — who would have otherwise contributed average or below average performance — substantially benefited from completing the rewriting task. One of my top performers (P22) explained that the task was helpful to her because it forced her to engage more deeply:

> It wasn't as apparent to me until I started having to go in and actually type it and reword the writing. So to me that was interesting because I'm like "oh my goodness, how many things do I read and just let go over my head?" And even working with my clients now I'm like "how many things have they told me about, but because I wasn't there to physically look at it and work through it, did I just let fall under the rug?" *It has me questioning a lot of things because I think reading it out loud and actually being able to go in the document and start taking things out and adding things helps bring a light and see things for what they truly are.* (P22)

[17] The Silhouette Coefficient [183] is a popular and useful way of evaluating the quality-of-fit of cluster data. In my case, the Silhouette Coefficient measures how close participants within a cluster are to one another compared to the participants in the next most similar cluster.

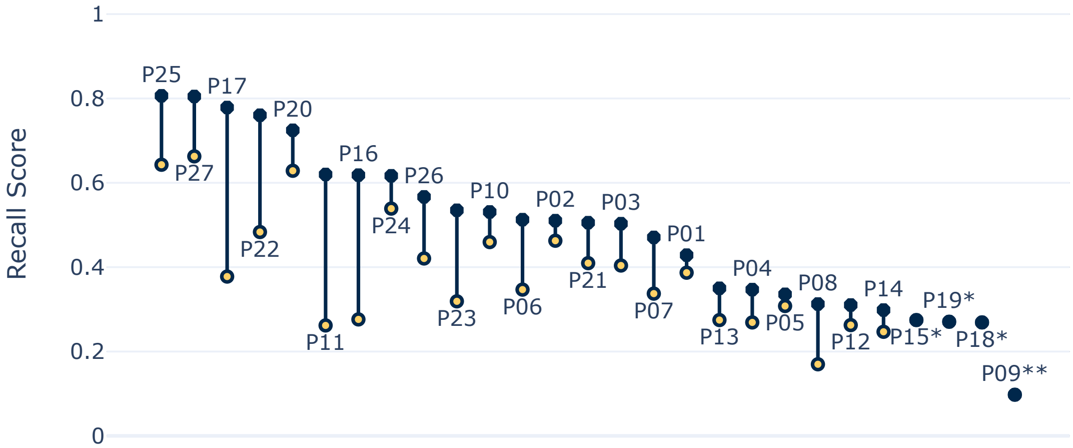


Figure 4.5: The average recall score by participant ordered from greatest to least[18]. Each vertical line shows the participant's average recall score after completing the highlight task on the bottom and their average recall score after completing both tasks on top.

Similarly, P23 added that the editing task played into her skill set as a writer, which also helped her engage more critically with the facts of the snippet rather than the emotional aspects:

> I'm a writer and I like to edit. I think attention to detail and *being able to take out the noise and reread it with a different critical eye is very useful in something like this*, versus just reading it. And even editing it with only emotion, I think when you put the emotion aside and you look at the facts, and you look at it critically, you can see what else may be wrong with this scenario. (P23)

In summary, I have examined my research questions by analyzing how my participants viewed the three text snippets and then conducting three follow-up analyses. In my analyses, I identified the issues that must be considered when constructing a Justice Panels system. In the next section, I will discuss the three main challenges I see in constructing a system like Justice Panels, based on my findings. I will also suggest potential solutions.

[18] Not all participants were able to complete the edit task due to the semi-structured interview running more than 1.5 hours. In total, this includes 6 participants in snippet 2 and 8 in snippet 3. Three of the four lowest-scoring participants shown in Figure 4.5 did not complete the edit task for snippet 2 or 3 (marked with "*"). Additionally, my lowest-scoring participant did not exhibit expertise in their evaluations (marked with "**"). I include them in my analysis for completeness, but I limit the use of their data when drawing conclusions from my findings.

4.5 Discussion

In this work, I set out to examine the expertise necessary to understand social references. I did so by exploring the challenges around creating *Justice Panels*, crowdsourced panels of social justice experts designed to find othering rhetoric. In my examination, I conducted a study of how 27 people who have expertise and interest in social justice identify racist and prejudiced words and phrases within three short text snippets derived from popular news articles that varied in the target marginalized community and subtlety of rhetoric. I conducted two analyses guided by two research questions: 1) the words, phrases, social references, dog whistles, stereotypes, and narratives that participants called attention to and determined how they build upon each other to form othering argument structures, and 2) the knowledge, skills, and experiences that participants pulled from to identify racist and prejudiced phrases.

In the first text snippet, I observed the least amount of variance in participant responses and the most obvious example of othering rhetoric (compared to the other three snippets). I observed that participants' social justice expertise was *additive*, where the more relevant lived experiences they had, the more of the phrases like 'illegal aliens' and 'from Guatemala' they could detect. I saw that the second text snippet had participant responses that varied more; the snippet was less additive. My participants saw othering in phrases such as 'paid for by taxpayers' as well as the combined use of the term 'refugee', and the phrase '$9 billion every five years.' I witnessed the most experienced participants identified most phrases, but to varying degrees, suggesting what I call *fragmentation* of the required expertise. In contrast, snippet 3 had the most variance in the responses. My participants disagreed about the impact and the contribution to othering for a number of phrases, including 'his mother was addicted to crack', 'poor, rundown, and violent', and 'South Side of Chicago'. I believe this disagreement reflected a combination of the differences in participant expertise and disagreements about the impact of phrases.

Based on the findings from these analyses, I conducted three follow-up analyses (see section 4.4.4) intended to provide further insight into my research questions. Specifically, I analyzed what was most helpful for participants as they completed my tasks, what made participants disagree with one another, and how my annotation tasks impacted participants' thought processes.

From these follow-up analyses, I came to three design conclusions that I believe are critical to consider when creating systems that recruit Justice Panels. For each design conclusion, I discuss their implications for developing a system, such as the challenges they introduce and potential solutions to those challenges. I explain each design conclusion below, and I conclude with a discussion of future work. See Table 4.2 for a summary of how my findings build on prior work.

Findings from Prior Literature	My Findings
— Enables expertise recruitment based on participant characteristics [82]	— Recognizes that crowd expertise is considerably complex and that there is a need to move toward multifaceted, contextualized, and situated views
— Recognizes the significance of disagreement in crowd responses [112] — Moves toward distinguishing between ambiguity and disagreement [39] — Identifies the significance of cases with high disagreement [92]	— Works toward identifying the sources of disagreement that result in differences in crowd responses
— Introduces the use of anchor comparison for disentangling social references from content [105]	— Extends the use of anchor comparison to engaging multiple participant skills for improving individual performance

Table 4.2: A summary of my findings compared to prior literature.

4.5.1 Design conclusion 1: social justice expertise is multifaceted, contextual, and situated

First, I argue that social justice expertise is multifaceted and contextually dependent on the content to be analyzed and the context around which that content is meant to be consumed. In my study (see Section 4.4.4.1), I found several categories of lived experience that helped determine participants' general expertise, but the specific lived experiences participants used within each category varied tremendously. For example, in snippet 2 I found that participants' experiences dealing with the immigration system were very important for identifying racist and prejudiced words and phrases based around refugees. However, these experiences with the immigration system were not based solely on their experience of their own immigration, but were often based on the trouble their family or clients had experienced.

This finding poses significant challenges for systems that aim to recruit relevant panel participants. In particular, systems that rely on singular factors such as identity or immigration status would likely result in a panel that lacks important perspectives, but may be perceived to have the top experts. For this reason, I stress that it is important to deeply engage with a potential panelist's lived experiences in the selection process rather than essentializing them based on a few attributes.

Instead of using singular factors (e.g., identity) alone in the panel selection process, I argue that an approach that assesses social justice expertise across many factors would be more appropriate. This approach would make it possible to recruit a portion of top performers — participants that scored in the top 20% of recall scores — while ensuring that all potential panelists are able to achieve an adequate baseline performance. I discussed evidence of this in section 4.4.4.1 where I found that roughly half of my top-performing participants were consistently top performers in all three of the text snippets. A well-developed social justice measure could make it possible to identify these participants, such as my second top-performer who, by his identity alone (e.g., White man), may not be considered with identity-based recruitment.

Part of my contribution is that I introduce an approach that can be used to create gold standards for problems that require social justice or similar expertise. I demonstrated the efficacy of my approach through an analysis of my participants' interviews; I believe that the resulting texts and annotations generated could be used to identify participants with relevant expertise for circumscribed situations. Potential participants could be asked to complete similar annotation tasks as my participants, and then be compared against my participants' findings to determine their score. A cutoff could be used to select participants who meet a minimum threshold. Additionally, I argue that my gold standard — and approach for generating gold standards — is a useful step toward developing a more generic social justice measure. As is suggested by my findings in section 4.4.4.1, a complete measure would likely need to include both an assessment of media literacy skills and relevant lived experiences. Importantly, I believe that this expanded view of social justice expertise builds on the work of Gordon et al. [82] by making it possible to select participants that are a particularly good fit for the content they are recruited to evaluate.

4.5.2 Design conclusion 2: interpreting sources of disagreement is difficult

Second, I argue that the findings of a Justice Panel may be difficult to interpret since disagreements among panelists can come from at least two sources. With my text snippets and participants, I observed disagreement — the least with snippet 1 and the most with snippet 3. However, as I discussed in section 4.4.4.2, I observed indications that the source of this disagreement varied by snippet, where disagreements in snippet 1 were primarily due to *expertise divergence* (differences in participant's expertise), but in snippet 3, both expertise divergence and *impact disagreements* occurred (disagreements about the impact of phrases despite having similar relevant lived experiences). Particularly with snippet 1 I observed that the number of racist and prejudiced phrases participants caught was generally proportional to the diversity of their lived experiences around objectification and being targeted with accusations. In contrast, I saw evidence in snippet 3 that some participants with similar lived experiences disagreed with one another. For example, two of my participants disagreed about whether the description of Caldwell's mother as 'addicted to crack' is used in a racist manner or whether it is simply used to express resentment about his mother, despite both participants having worked as social workers and having parents struggling with drug abuse.

Determining the source of disagreement is important to understand the significance of Justice Panel results. While the majority opinion could be used to find the most obvious components of othering rhetoric, analyzing subgroup opinions could help identify more subtle components that require niche expertise to understand. I demonstrated the potential of this approach in snippet 1 where simple hierarchical clustering revealed an additive expertise structure, where specific phrases were tied to specific expertise. However, using the same approach is not as effective in determining the significance behind subgroups in snippet 3, since the meaning of subgroups was muddled by impact disagreements.

Understanding the sources of disagreements is also important because, as briefly mentioned in section 4.3.2.3, impact disagreements can muddle potential measures a system designer might use to find top-performing panelists. For example, the recall score I use in this study to tentatively evaluate participant performance does not adequately capture some expert decisions *not* to annotate specific phrases. While I was able to identify some decisions through the analysis of participants' interviews, it poses a problem for future systems that rely on similar quantitative measures.

It may be possible to uncover early indications that the source of disagreement among participants is complex by analyzing the structure of annotation data. I demonstrated this possibility in section 4.4.4.2 by examining the distribution of participants' recall scores (the percentage of phrases an individual participant caught out of all the phrases any participant caught). I think this approach is promising for determining the quantity of disagreement within a particular text snippet. Additionally, I used a measure called the *Silhouette Coefficient* (defined in footnote 17) to quantify the degree to which participants can be neatly grouped together into clusters based on the phrases they find. I argue that in cases where participants can be neatly grouped together, it is likely that there are relatively few sources of disagreement and that further analysis could potentially determine a specific source (e.g., expert divergence or impact disagreements). In cases where participants cannot be neatly grouped, I argue that there are likely multiple sources of disagreement at play and that more data may be necessary to understand the significance of subgroups.

This approach extends prior work in interpreting crowd worker responses by introducing an early indicator that can be used to determine appropriate follow-up analysis. For example, my approach could be used to extend the capabilities of Kairam and Heer's system [112] by helping determine the significance (or lack thereof) of participant clusters. In cases where there are relatively few sources of disagreement, it may be possible to emulate a part of my interview process by issuing follow-up tasks to participants within relevant clusters asking about why they did or didn't think specific phrases were worth annotating. The result of this follow-up task could help determine the significance of specific participant clusters. Likewise, my approach extends Gurari and Grauman's CrowdVerge system [92] by helping decide which cases have many sources of disagreement and which could benefit from gathering more panelists' perspectives to help disentangle. While CrowdVerge can be used to identify cases where there *is disagreement*, my approach could help identify the general *amount of sources of disagreement* — an important distinction in determining whether follow-up data collection is likely to be useful.

4.5.3 Design conclusion 3: anchor comparison is surprisingly useful

Finally, I argue that the anchor comparison approach I used in my second annotation task (editing task) can be used to improve individual performance. While the approach was originally used to mitigate the conflation of content with social references [105], my study shows that it can also be used to engage participants with a broader skill set. In my third follow-up study in section 4.4.4.3 I found substantial evidence that completing both of my tasks helped nearly all participants identify phrases they were not able to identify when completing the highlight task alone. I believe that this is an important finding that should be considered when constructing annotation tasks so that panelists can best engage with the content.

To ensure that participants are able to reach a consistent depth of analysis, I suggest using multiple rounds of annotation including identification and editing tasks. In particular, I find that my approach extends my findings from chapter 3 by applying anchor comparison to a new context that boosts the performance of nearly all participants, including a substantial increase in some that moved them from average performers to top performers. I found evidence in my interview data that suggests this was in part due to the editing task making use of a different skill set than what was used in the highlighting task. Specifically, a few participants suggested that the editing task enabled them to think more critically about how the facts of the snippet contribute to othering rhetoric, rather than focusing on the emotional elements alone. My suggestion demonstrates an important application of the anchor comparison approach that should be used to improve the individual performances of Justice Panel participants.

4.5.4 System level implications and future work

I have taken important steps toward building systems that can recruit Justice Panels for identifying encoded othering rhetoric. I believe that these systems may have the capability of identifying emerging narrative archetypes that seek to other people-groups and combat them before they are made popular in the public. Importantly, while I have found a few promising directions for identifying many forms of expertise, I find it unlikely that systems will be capable of consistently recruiting top experts for panels. However, I believe there is still hope for future research.

As is implied by my first and second design conclusions, there is significant need for further research into developing measures for social justice expertise and for distinguishing when specific disagreements can be attributed to a lack of expertise. I have made steps in this direction by contributing an approach for generating gold standards, a possible gold standard that could be used to identify social justice aptitude for circumscribed situations, and by contributing an approach that may be capable of identifying cases where there are relatively few sources of disagreement. While I believe these metrics will be valuable and effective, they still need evaluation for efficacy, and there remains a need for approaches that engage more thoroughly with the situated nature of social justice expertise and can precisely identify sources of disagreement.

Likewise, while I found that anchor comparison was helpful for all of my participants and substantially helpful for some, there is a need to explore additional tools for engaging panelists. Similar to how I found that anchor comparison did a better job in making use of participant's journalism skills (see section 4.4.4.3), it is possible that future tools could make use of additional skills panelists offer. I hope that my work will inspire additional exploration with potential tools.

4.6 Conclusion

In this chapter, I have studied the aspects of crowdsourced expertise and explored the challenges associated with building systems that recruit *Justice Panels* — crowdsourced panels of social justice experts — for identifying othering rhetoric. In my exploration, I studied 27 people who have substantial expertise and passion for social justice to answer two research questions: 1) what do social justice experts see as othering rhetoric? and 2) what are the skills, knowledge, and experiences social justice experts use to identify othering rhetoric? I asked participants to explain their thought processes as they completed two annotation tasks, first asking them to highlight racist and prejudiced words and phrases, and second to edit the text to remove such phrases. My participants identified many phrases that, while not in constituting racist or prejudiced language in and of themselves, helped create othering rhetoric by acting as *semiotic payloads* — smuggling connotations through devices such as narratives, stereotypes, tropes, dog whistles, and social references.

Upon answering these questions, I arrived at three design conclusions that I believe are important for the direction of future work: 1) that social justice expertise is multifaceted, contextual, and situated, implying that it cannot be reduced into a small collection of demographic factors, 2) that there are at least two sources of disagreements among participants including *expertise divergence* and *impact disagreements*, and 3) that *anchor comparison* is a surprisingly useful technique for improving participants individual performance. I believe these three conclusions are important contributions not just for designing Justice Panel systems, but also for crowdsourcing and human-AI systems that rely on the expertise of diverse annotators. I hope my findings will motivate the development of systems that engage more deeply with panelists' skills, knowledge, and experiences. In the next chapter, I will propose an alternative approach for ensuring social reference processing systems have adequate context to understand social references.

CHAPTER 5

Guided Context Building: Supporting Context Building Effort Around Social References

In this study, I examine how the process of building context around social references can be augmented with AI support. I do this within a problem domain where human-AI collaborations are rapidly emerging: language translation. This problem domain has the advantage of being well understood, including potential evaluative metrics and a well-documented process where people research terms for the connotations they give off. Learning how to design AI support systems in this domain helps me understand how AI support could be designed in other professional social reference processing domains.

More specifically, I examine the *context building* that professional freelance translators (PFTs) undergo, in which they gather the connotations around terms to understand their utility within the translation context. I argue that a context building process that combines humans and AI suggestions is an alternative approach for identifying relevant context around social references that could be applied in the other domains that I mention in this dissertation, including for identifying manipulative language and othering rhetoric. I examine ways to support this process by first conducting a contextual inquiry with PFTs in which I uncover the ways in which they engage in the process of context building. Based on the findings, I then propose and develop an approach I call *guided context building* that supports the context building process by directing effort toward places where it is known to be necessary. I evaluate this system, implemented on top of a commercial computer-assisted translation (CAT) tool, in a user study of 98 PFTs where I demonstrate its feasibility for shaping context building effort and its impact on task outcomes. My work demonstrates that providing support for context building within collaborative human-AI social reference processing systems can be best accomplished not just by improving the quality of AI suggestions, but also by providing collaborative tooling that helps people determine when to evaluate those suggestions.

5.1 Introduction

Building an understanding of the connotations elicited by terms is an important part of doing effective social reference processing. In particular, the process involved with gathering context surrounding social references, a process I refer to as *context building*, is essential for ensuring that a system has all of the relevant information needed to understand them. I explore approaches for supporting this process in order to improve the capabilities of social reference processing systems.

Specifically, I build upon the work of the previous two chapters (3 and 4) by exploring how artificial intelligence (AI) can be used to aid collaborations between people by supporting the context building process. Such an approach could make it possible to compliment my prior two approaches by suggesting possibly relevant surrounding context for social references to an expert panel, who could then leverage their expertise to determine their relevancy. This would be useful since, as I discussed in chapter 4, it is extremely challenging to identify top experts. However, it may be more feasible to recruit a panel consisting of people with adequate expertise, and then bolster their expertise by filling in knowledge gaps with AI support.

In this chapter, I learn about how to best provide AI context building support by studying a well-established commercialized domain where human-AI context building collaborations are already common: language translation. Language translation offers a unique opportunity to learn how professional freelance translators (PFTs)[1] collaborate with neural machine translation (NMT) [16] systems while undergoing the process of context building to understand the connotations surrounding terms they are translating. Learning how to provide AI support for context building in this domain enables me to draw insights into how AI support can be provided in my other social reference processing domains.

With the introduction of NMT, the use of human-AI systems within modern language translation workflows is now commonplace [58, 113]. Despite recent advances, publication-quality language translations are still largely written by professional translators. However, the use of NMT systems in translation work is increasing rapidly [54], and professional freelance translators (PFTs) can achieve higher quality and higher speed by engaging in human-AI collaboration with NMT systems [88, 87, 102]. Part of my goal is to understand how these collaborations currently take place and how AI influences the context building process.

Through my exploration of human-AI language translation systems, I provide field-based evidence of the role AI could play in supporting context building for social reference processing systems. In the domain of language translation, I find that PFTs go through context building to develop an understanding of the meanings, connotations, and applicability behind suggestions made

[1] Throughout this chapter, I use the acronym "PFT" instead of "translator" to avoid confusion between professional *human translators* and a *machine translator*. I believe the term adds greater weight to the humanity and expertise that is required to do professional translation work.

by the NMT system. I argue that context building is a substantial component of the human effort in a human-AI translation collaboration, and that the process can be shaped through tooling to influence both the context building process as well as translation speed and quality.

Through a series of semi-structured contextual inquiry interviews with PFTs, I find that context building for translation tasks involves gathering connotations of both terms and potential translations. I identify an inventory of different modes of context building and catalog some of the typical activities involved (e.g., web searches) and sources consulted (e.g., bilingual dictionaries). Critically, I find that AI-generated translation suggestions play a central role in context building, indicating that external research and human-AI collaboration are closely coupled components of a modern language translation workflow.

Based on these interviews, I propose an approach for shaping the way that PFTs evaluate AI-generated suggestions that I call *guided context building*, in which I direct their context building effort by populating a panel of contextual information that is relevant to the terms in a document that require the most context building. I experiment with two forms of guided context building: automatically curating context from popular dictionaries and concordances, and simulating collaboration between PFTs (targeted guidance). I implement this approach within a commercial computer-assisted translation (CAT) tool and evaluate its impact with a user study of 98 PFTs.

My results show that is feasible to use guided context building to shape overall context building effort. Showing contextual information in the panel without targeted feedback increases context building activity, whereas including targeted guidance focuses context building effort toward specific terms. In both cases, I find a shift in the information that is sought by PFTs and that additional context building increases total translation time. In some cases, additional context building time is also correlated with increased translation quality of the specific terms being researched.

Based on my evaluation from translation tasks, I conclude that providing AI support for context building within collaborative human-AI social reference processing systems involves not just improving the quality of AI suggestions, but also by providing collaborative tooling that helps people determine when to evaluate those suggestions.

In this chapter, I make the following main contribution:

- I show the importance of building tooling that supports the human effort required to evaluate AI-suggested context within collaborative human-AI social reference processing systems.

I accomplish this by making the following additional contributions:

- Uncovering how professional freelance translators (PFTs) collaborate within a human-AI system by undergoing the process of *context building* — which involves gathering relevant context surrounding social references. I provide evidence that PFTs undergo this process in part to evaluate the utility of MT suggestions.

- Proposing the idea of *guided context building*, where context building effort is directed toward examining specific social references, connotations, or resources. I implement this guided context building approach on top of an existing commercial CAT system.
- Evaluating my approach through a user study with 98 PFTs and demonstrating its feasibility for shaping PFTs' context building effort.

5.2 Related Work

I build upon the work of prior researchers in AI, HCI, CSCW, and the social sciences in understanding PFT's relationship with AI and how to best support their collaboration. Among the vast literature at the intersection of these fields, the most relevant topics include translation practices, use of AI within translation workflows, and human-AI collaboration.

5.2.1 Translation practices

Long before the widespread use of AI, social scientists have spent significant research effort cataloging a variety of practices that professional translators at language service providers (LSPs) engage in. Translation is largely understood as a highly collaborative [58] social process [151] in which the translator has the agency to shape the meaning of the original source document [118]. Professional translators make use of a large variety of resources that have, like in many other domains, transitioned from mostly physical [180] to now mostly online [179]. Common resources include dictionaries, concordances, and CAT software that feature termbases[2] and translation memories[2] [55]. Extensive prior ethnographic research in translation workplaces has demonstrated that professional translators are hesitant to adopt new technologies [124, 125], but through numerous improvements to the ergonomics and usability of these resources, they now frequently make use of them [203, 207]. AI has the potential to serve professional translators' needs by acting as a catalyst for dynamic collaboration, but many still view the technology as a threat to their agency [113].

5.2.2 AI within translation workflows

My work comes during a significant shift in translation workflows as a result of an influx in demand for translation work and of CAT tools that incorporate AI. While the bulk of translation work has historically been completed in-house, increased connectivity due to globalization and the explosion of content on the internet has led to a rise in demand for faster translation work with lower

[2] Databases that track and suggest prior translations of terms and similar source text to aid PFTs with maintaining consistency in their projects.

demands on quality [133, 223]. PFTs are naturally good candidates for such work since they are able to provide faster turn-arounds due to being available on-demand [71]. Experience requirements vary widely between translation jobs, including jobs that require professionals with extensive experience [223], jobs that make use of crowd workers with no professional experience [9], jobs that make use of monolingual crowd workers [99, 100], and jobs that rely on MT alone [210]. For this reason, researchers have sought to build translation workflows that enable more efficient translation while enabling PFTs to tailor the translation to the client's needs [26, 167].

5.2.2.1 Post-editing workflows

The first of these translation workflows to emerge to the market is the *post-editing workflow* that is now present in many popular systems such as CASMACAT [5], Matecat [66], MemoQ [96], and SDL Trados [97]. CAT tools that leverage a post-editing workflow pre-translate text with MT and then task PFTs with correcting the translation after the fact [7], an approach that improves translation efficiency by reducing PFT's cognitive load when getting started with a translation [88]. While post-editing is more effective than unassisted translation in contexts where the MT is accurate, further research has raised concerns about its efficacy in settings where the MT is prone to errors [57, 122] which require labor-intensive edits to correct [90, 159].

5.2.2.2 Interactive workflows

To address these challenges, recent CAT tools such as Lilt [95] and Transmart [102] make use of a less restrictive workflow — called the *interactive workflow* — where the MT provides realtime translation suggestions to the PFT in an auto-complete style [164]. The interactive workflow improves upon post-editing by providing PFTs with more agency in their translation, leading many to prefer it [119]. However, as Moorkens et al. report, the utility of the interactive workflow is still bound by the quality of the underlying MT, which many PFTs still distrust and are hesitant to make use of [160]. Coppers et al. developed the Intellingo system to explore how different extensions to an interactive interface help with intelligibility [48] and found that intelligibility features are preferred by PFTs when the information presented isn't already in a PFT's available knowledge.

5.2.3 Human-AI collaboration

My work builds on a foundation of research on human-AI collaboration within the HCI and CSCW communities. While human-AI collaborations have long been discussed in the context of automation [126], they are now also widely discussed in the context of decision support tools [10, 134, 219], within crowd-based systems [42, 149], and in a large variety of other contexts such as recommending movies [19] and music [146]. However, in nearly all domains, the role of the concepts of *interpretability*, *trust*, and *reliance* are frequently discussed as prime factors in supporting effective human-AI collaborations. In particular, an AI's interpretability is known to impact peoples' ability to predict its actions [34, 174], a person's trust in AI is known to impact their reliance on the AI [77], and a person's reliance on AI is known to impact many sociotechnical outcomes [85, 86].

5.2.3.1 Trust in and reliance on AI

Use of AI assistance in human-AI collaboration is importantly governed by trust and reliance relations. The relationship between trust and reliance has become of particular interest to many AI practitioners who have long sought to model how people come to trust an AI in hopes to better calibrate their reliance [126, 163]. Trust is differentiated from reliance (or adherence) in that it refers to a person's *beliefs* about the AI whereas reliance refers to *realized behaviors* [50]. More recent research has expanded on these initial models by incorporating the concept of *trustworthiness* [108] (whether trust in an AI is warranted) and by identifying a number of social factors that influence a person's trust in AI as compared to groups of people [127]. Others have examined more specific contexts that people calibrate their reliance on AI such as when performance feedback is limited [135], when determining to give out child welfare [115], and when deciding to use machine-generated code within critical control system software at NASA [211].

5.2.3.2 Calibrating reliance on AI

Within much of this research, many have raised concerns over the potential that people may improperly calibrate their reliance on AI. In particular, after introducing AI in various settings, prior work has found that it may lead clinicians to take less agency when taking medical annotations [130], may lead people to exhibit increased bias [85], and may lead to increases in perceived risk in pretrial assessments and more risk-adverse decisions when deciding whether to give out government aid [86].

To combat these concerns, a line of research has developed that aims to help calibrate people's reliance on AI. Common strategies include providing explanations and confidence scores for AI recommendations [226] which, while leading to improvements in overall human-AI team performance, can also backfire by increasing overreliance on the AI [17]. Follow up work demonstrated that this effect can be mitigated with cognitive forcing interventions that compel people to engage more thoughtfully with AI-generated explanations [31]. Despite these developments, reliance on AI is still largely treated as a binary where the human either accepts or rejects AI suggestions.

In my work, I examine a case study of how this assumption fails by demonstrating that reliance may be more accurately understood in terms of *effort* and I extend the literature by introducing an approach for shaping the effort people extend toward relying on AI. In addition, I address a growing desire to see research on PFTs' functional relationship with AI and AI's role within broader sociotechnical translation systems [117, 184]. I provide insight into this topic by uncovering how PFTs assess the quality of AI suggestions and by designing an approach that assists PFTs in their evaluation of AI. In the next section, I discuss the first steps I take by conducting a contextual inquiry with PFTs.

5.3 Prestudy: Contextual Inquiry with Professional Freelance Translators

In this section, I will discuss the first steps I took to understand how PFTs collaborate with AI within their translation workflows to understand the connotations surrounding terms. Namely, I conducted a contextual inquiry with professional freelance translators to build an understanding of the processes and resources PFTs use in their daily translation work, and I used this information to design my system and inform my main study. I will first discuss the details of my method and then I will walk through my main findings.

5.3.1 Method

I conducted six semi-structured interviews with professional French freelance translators in the form of a contextual inquiry [213]. I used a contextual inquiry because it enabled me to better examine the thought process and work behaviors undertaken by the PFTs within their typical work environment. I recruited participants who had demonstrated a strong track record working with Lilt and had substantial experience working with other freelance translation agencies, compensating them with an hourly rate according to their negotiated contract with Lilt.

Within the interviews, I provided my PFTs with a state-of-the-art NMT system[3] and asked them to translate three text snippets that range in both the domain and difficulty. Specifically, my text snippets are two sentences in length and cover topics such as a corporate code of conduct, marketing for an Asics shoe, and a patent for a locomotion assisting device (see Figures 5.1, 5.2, and 5.3 for the exact snippets and their context). At the beginning of each translation task, I provided PFTs with the full text and I pointed them to the portion to be translated. While my participants translated each text snippet, I probed them for their thoughts and asked them to explain their decision-making process. With consent from my participants, I recorded and transcribed each interview. I analyzed

[3] I partnered with the language service provider (LSP) Lilt throughout this study, including by using their state-of-the-art translation system. The system provides PFTs with MT suggestions in an interactive workflow fashion (see Figure 5.4). All interview participants had extensive experience working with the system prior to their interviews.

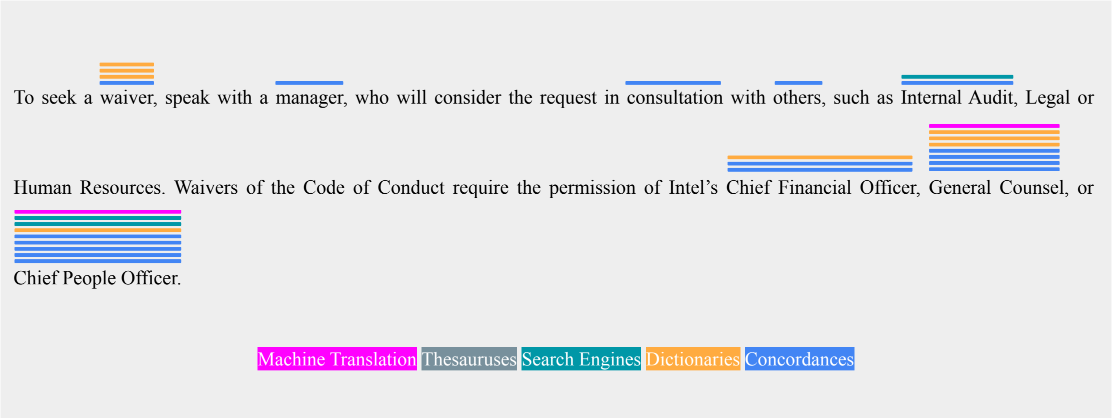


Figure 5.1: The resources participants in my initial study used to build context for various terms in the code of conduct text snippet. Individual lines represent cases where a PFT used a resource to search a term.

my data using Clarke’s Situational Analysis [45], an updated form of grounded theory [200] that recognizes the importance of Symbolic Interactionalism within interpretivist data analysis.

5.3.2 Results

Within all six of my interviews, I found that participants spend a substantial portion of their time undergoing a process I call *context building*[4] where PFTs research the connotations around source language and potential translations. In all three of the text snippets, I observed that nearly all context building effort was centered around a few terms. Participants used a variety of resources to build context for each of these terms including dictionaries, concordances, thesauruses, search engines, MT-based tools, and many other unstructured resources. I will discuss the details of my interview findings by walking through each text snippet, offering a quote from the participants where appropriate. (Note: All participant quote data presented here has been anonymized and edited for clarity, and my emphasis has been added in italics.)

5.3.2.1 Code of Conduct

My first text snippet consists of two sentences from a legal document about the procedures for waiving a provision from Intel’s code of conduct. During the interviews, I provided participants with the entire document and pointed them to the portion to be translated. The language is the simplest of the three text snippets: the most difficult terms consist of the names of various roles

[4] I assign a name to a common process PFTs undergo. Karamanis et al. describe some of the practices PFTs at a language service provider use while completing this process prior to the widespread adoption of MT within translation workflows [113]. Now that MT usage is more common, I examine how MT influences PFT’s use of context building.

within Intel's management. My participants focused their context building effort on five terms including "waiver," "Internal Audit," "Chief Financial Officer," "General Counsel," and "Chief People Officer" (see Figure 5.1).

In the first of these terms, "waiver," participants used multiple dictionaries in their context building effort. Specifically, participants leveraged the official dictionary produced by the Quebec government called *Grand Dictionnaire Terminologique* (GDT), a well-known dictionary in France called *Larousse*, and a dictionary operated by the Canadian federal government called *TERMIUM Plus*. Given that there are many potential translations for the term in French, participants used these dictionaries to find an option that fits the translation context and is used both in Canada and in France. Most participants kept the MT suggestion in mind while they made their search:

> I'm going to look up waiver just to be sure. *I know this suggestion here is good, but since this is like a document that as some sort of people value, I want to be extra sure.* (P4)

Despite participants acknowledging potentially better translations for the term, most participants ended up using the MT suggestion anyway. P3 explained that they preferred to keep MT suggestion because it can help maintain consistency with other PFTs working on the same project:

> [The MT suggestion] would be acceptable in a legal context, I will leave it as is. I will leave it as is even if I'm thinking that I might have a prettier word for it. Intel is a huge contract, so there's probably like 10 [PFTs] working on it at one time on different sections or whatnot. Therefore I'm going to leave what the platform suggests [...] *I have found that when more than one [PFT] works on the same contract it is best to keep the suggestion from the platform [...] to maintain consistency across multiple people that might be working on the same client at the same time.* (P3)

In contrast, participants made mostly unstructured searches and used concordances in their context building for the term "Chief People Officer." Nearly all participants had not seen the term used before and many suggested that it may be an alternative title for the head of Human Resources. For this reason, most participants started their search with the concordance *Linguee* where they would search for other corporate translations of the title:

> So because [the title is] made up, I'm gonna go in Linguee. *I'm seeing here that McDonalds also uses this term, and that in Canada, they call it "chef des Ressources humaines", so I'm going to go with that.* (P4)

For some participants, context building did not end after the initial search. Instead, they continued their effort by focusing on their proposed translation. For example, P4 sought to verify

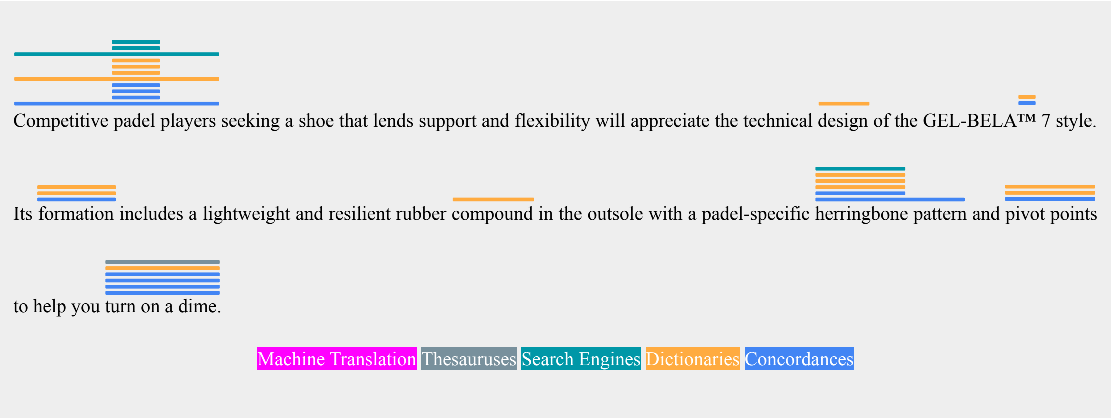


Figure 5.2: The resources participants in my initial study used to build context for various terms in the marketing text snippet.

their initial translation by looking up Intel's corporate roles on LinkedIn to see if the 'Chief People Officer' differs from the head of Human Resources. Likewise, P3 continued by searching the term in their secondary language, Spanish, to see if Human Resources is used in Spanish as well:

> If they're calling it 'Chief People Officer' even in Spanish or if they're calling it like [another Spanish translation] or something to see if they're translating it or if they're keeping it 'Chief People Officer'. *I'm looking to see if they've decided that 'Chief People Officer' was going to be 'Chief People Officer' in all of their languages.* (P3)

5.3.2.2 Marketing

My second text snippet consists of two sentences from the product details of an online listing for an Asics shoe intended for playing the sport padel[5]. Again, during the interview, I provided participants with a link to the product listing and I pointed to them to the portion to be translated. In general, participants found the main terms in the text snippet to be more challenging to translate than in my first text snippet, and they focused their context building effort on the terms "padel," "formation," "herringbone," "pivot points," and "turn on a dime" (see Figure 5.2).

For the first term, "padel," participants used a combination of condordance, unstructured searches, and dictionaries in their context building efforts. Most participants were unsure of what the term is referring to, so they generally started by searching the term on Linguee, Google, or Wikipedia. After learning about the term, some participants moved to dictionaries such as GDT or Termium to check for an official translation. Coming up short of an official translation, all participants stuck with the MT suggestion, which matched their findings from their initial search. P3

[5] Not to be confused with *paddle tennis*. See https://en.wikipedia.org/wiki/Padel_(sport)

explains their thought process as they moved between these resources:

> So the target in this case is people that do the sport. If it were a snowboarding document or snowboarding website, there are words because the sport is a little bit older than padel. [...] If I'm writing a Sports Medicine article about snowboarding and I'm translating that into French, I will go look for the currently approved French words by the Language Bureau. If I am translating a snowboarding document that is for a post on Instagram, Reddit, Twitter, Facebook, and whatnot, then *I might not use the "formal" French word, but more the lingo that the target recognizes*. It's like, longitudinal versus lengthwise, it means the same thing. Who am I writing to? Who's going to read this? (P3)

As in the previous text snippet, participants often kept the MT suggestion in mind during the context building process. For example, upon finishing their context building efforts for the term, P6 weighed the evidence to determine whether the MT suggestion is appropriate in the context:

> Looking at what [the MT] has proposed and looking and thinking about what I would say, I would keep what [the MT] has offered. *I'm looking at what they've written and everything looks like something you might say in French and what you would see in a regular French advertisement*. (P6)

The final term in the document, "turn on a dime," also saw a high frequency of concordance searches. However, unlike in their context building for "padel," participants used concordances for more creative purposes. For example, P5 describes how they used Linguee as a means to find new ways to translate the phrase:

> So for this part 'help you turn on a dime', I know what it means, but I also know that it's some kind of figure of speech and I'm gonna need a little bit of help in order for it to make sense in French. [...] So here is the problem, there is no actual word in French for that. [...] Sometimes you're focusing on some words way too much and you cannot think of synonyms or other ways to tell the same thing. So Linguee is like Google Translate, it's not the best thing ever. If you want to translate a full sentence, that's going to be hectic. But *if you actually want to have different ways of putting the same word in certain contexts*, it's gonna be really useful — especially in French. (P5)

5.3.2.3 Patent

My third and final text snippet consists of two sentences out of the abstract of a medical patent for a locomotion assisting device [78]. Participants were provided a link to the full patent and the

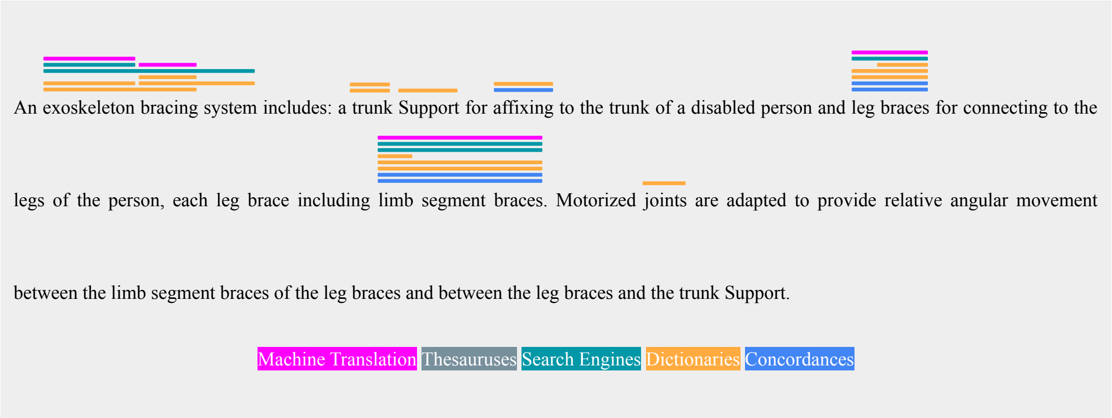

An exoskeleton bracing system includes: a trunk Support for affixing to the trunk of a disabled person and leg braces for connecting to the legs of the person, each leg brace including limb segment braces. Motorized joints are adapted to provide relative angular movement between the limb segment braces of the leg braces and between the leg braces and the trunk Support.

Figure 5.3: The resources participants in my initial study used to build context for various terms in the patent text snippet.

location of the portion to be translated before they began translating. The text snippet contains the most technical language of the three snippets, and participants found it to be the most challenging to translate. Like the previous snippets, participants focused their context building efforts on a few of the terms including "exoskeleton bracing system," "trunk," "affixing," "leg braces," and "limb segment braces" (see Figure 5.3). I will again discuss a few of them.

For the first of these terms, "exoskeleton bracing system," about half of my participants started their context building effort with a dictionary search and another half with a Google or MT search. Almost all participants mentioned that the domain (medical patent) impacted how they went about building context for the term. Specifically, participants mentioned the need to find translations from trustworthy resources. For example, P6 discussed how they switched from using their go-to resource to using the official terminology base of the European Union called *Interactive Terminology for Europe* (IATE) that they believe is better suited for the domain:

> For this translation, *I think I'm going to use a different website just because I see there's some very technical and medical vocabulary*. The website I'm going to use has more technical terminology, I think I can find more answers that I'm looking for here. [IATE is] nice because with this standard terminology in specific domains, not only do you have the translation, but you can see where the word came from, just to know that it's an actual word that's used and exists. You could go [into the results] and see the term reference and a definition, and so *I think it's better in this type of medical context*. (P6)

Likewise, P4 described a similar thought process as they used the dictionary GDT. They explained how they came to the conclusion that their translation for the term fits the medical domain:

> Here I can file by area of interest, and I want medicine. And so [the first definition] is "the action of applying a device to the body or part of the human body in order to replace them, support them, or supplement their function". So this is clearly what we're talking about here with this exoskeleton. So this is the right word. And I'm not gonna look anywhere else because the way this is described, it's the perfect word. So I'm just gonna go with that. Also, *because I know how this database was created. They source things, it's terminologists that built it. So I'm confident in the quality of the terms that are on there*. (P4)

Again, like in prior text snippets, participants followed up their context building by returning to the MT suggestion and considering whether it was applicable. For example, P2 considered using the MT suggestion, but ultimately rejected it after finding a more specific term:

> I just replaced it because [the MT suggestion] is not the correct word for "bracing". [The MT suggestion] here would be to nail something on the wall, to fix with screws or glue or whatever. You want something that's more holding together, right? It's not the same thing. [...] That's the thing, *when you're getting into very narrow fields like medical law, things like that — anything that's got to do with the Navy or the Marines — they have very specific terms like "aft" or "starboard" or "port" that you don't use otherwise in normal life, you use different terms for that*. (P2)

In the last term of the text snippet, "limb segment braces," participants started their context building in a similar way as they did for the first term. However, unlike the first term, participants came up short of results from official resources, so some turned to unofficial ones (e.g., Linguee searches) only to arrive in a similar position. Short of findings from relevant resources, participants described how they effectively "made-up" a translation for the term that they thought was appropriate for the context. Most participants checked their proposed translation by searching it online, like P4 who searched their translation on Google:

> To be sure that [my proposed translation is] a thing that people have said in the past, you can just look it up [...] I had no results for the combination of words I used. No one has ever combined these words before in that way, which means I'm wrong. Unless no one has ever used the English combination of words in that way either. (P4)

P4 continued by searching the English term in Google, which also returned few results:

> No one had also combined these words that way in English, so it makes sense. So I'm still confident in my translation after I've verified that, because that's the thing that happens [when translating] is you come across a combination of words that no one ever used before, so I'm going to stay with that. (P4)

Mode	Explanation	Interview example
Searching for an "official" translation	Looking at authoritative sources for definitions or formal translations of a term.	Using GDT to find a government recommended translation for 'waive'.
Searching for in-group lingo	Looking for a translation of a term that is used within a particular subgroup or community.	Looking up the French Wiki page for 'padel' to find the community translation.
Searching for credibility signals	Looking for reasons to accept or reject a particular translation of a term.	Googling 'limb segment braces' to see if it returns results to be sure it is a niche term.
Brainstorming	Enumerating potential translations of a term to help find a more creative translation.	Using Linguee to find translation ideas for the term 'turn on a dime'.

Table 5.1: I found four modes in which PFTs engage in context building

My participants made use of multiple resources and underwent numerous thought processes during their context building efforts. As I have described, participants used these resources to build an understanding around terms and the connotations of their potential translations. Participants interacted with MT throughout this process, and they often focused on whether they should accept or reject the MT suggestions. In the next section, I will summarize my findings and how they influence the design of my system.

5.3.3 Discussion

Through my initial contextual inquiry, I first found that PFTs engage in the process of context building in a complex combination of ways and for a variety of purposes. Second, I found that they undergo this process in part to evaluate the utility of MT suggestions for that translation context. I will discuss each finding in more detail.

I have observed how context building can be a complex process that requires a deep social understanding of the source text, the client's needs, and the expected audience of the translated text. In the absence of direct communication channels that are common at many in-house LSPs [113], PFTs turn to various online resources instead.

In particular, I have observed PFTs engage in four different *modes of context building*[6] (see Table 5.1): searching for an "official" translation; searching for in-group lingo; searching for credibility signals; and, brainstorming translations. PFTs engaged in these modes to accomplish different goals such as searching for a known acceptable translation for a term due to it being backed by

[6] There may exist more than four modes, but I can only speak to the four I have observed.

authoritative sources, or searching for a translation of a term they think is more likely to be recognized within the target community. PFTs constantly swapped among these modes and many times even engaged in multiple at once. For example, for the term "padel" in the marketing text snippet, I witnessed participants start their search with Wikipedia searches (in-group lingo), then swap to searching formal dictionaries like Termium and GDT (official translation), all while accumulating evidence that the MT suggestion is applicable in the context (credibility signals).

MT significantly influences the way PFTs engage in context building. On a few occasions, participants explicitly mentioned a preference to keep the MT suggestion given it is tolerable in the context. As participants explained, translation teams are made up of PFTs who often don't know each other and may be located in time zones across the world. Directly communicating with others is difficult and time consuming, but MT can help provide consistency despite the low communication environment by providing a default translation option. As a result, many participants put in additional context building effort toward building credibility around MT suggestions.

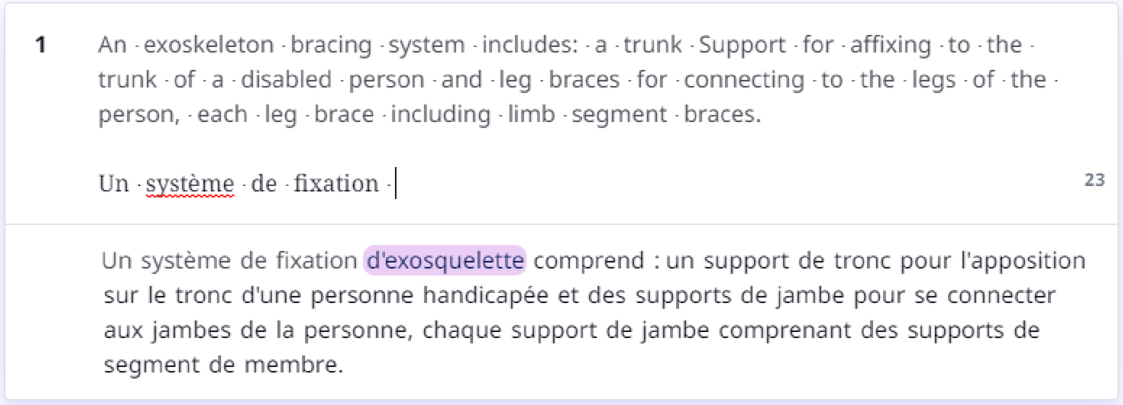
5.3.4 Design Goals

The findings from my contextual inquiry can be used to design a system that supports PFT-AI collaboration. Specifically, they suggest following design goals:

G1. Provide support for evaluating AI assistance — specifically around AI suggestions.
My prestudy found that PFTs based a substantial portion of their context building effort on searching for credibility signals within MT suggestions. While they did so for multiple reasons, I found that a major one was to maintain consistency with other PFTs. I aimed to augment this use case by providing PFTs with tools to evaluate the utility of MT suggestions.

G2. Supply credibility signals from resources that are known to be trustworthy.
In three of the modes of context building, participants relied on the credibility of external resources to evaluate the applicability of translations. In the case of "searching for the official translation", they used trustworthy resources to begin their search. In the other two cases ("searching for in-group lingo" and "searching for credibility signals") they often searched for relevant information first, but then evaluated the quality of that information based on how trustworthy they consider the resource. I sought to support these existing evaluation behaviors by validating potential translations with the credibility of resources that are known to be trustworthy.

G3. Accommodate the conditions of crowd-based labor.
As is the case with my partner LSP (Lilt), freelance translation has an increasing number of aspects in common with crowd labor. In particular, freelance translators are not expected to



1 An ·exoskeleton ·bracing ·system ·includes: ·a ·trunk ·Support ·for ·affixing ·to ·the ·trunk ·of ·a ·disabled ·person ·and ·leg ·braces ·for ·connecting ·to ·the ·legs ·of ·the ·person, ·each ·leg ·brace ·including ·limb ·segment ·braces.

Un ·système ·de ·fixation ·

Un système de fixation d'exosquelette comprend : un support de tronc pour l'apposition sur le tronc d'une personne handicapée et des supports de jambe pour se connecter aux jambes de la personne, chaque support de jambe comprenant des supports de segment de membre.

Figure 5.4: My CAT tool enables the PFT to interact with machine translation suggestions, which automatically update based on the text typed. PFTs can auto-complete suggestions with a hotkey.

stay with clients for extended periods of time, often resulting in high turn-over within translation projects. Additionally, it is increasingly common for LSPs to make use of open calls to recruit PFTs, which is a fundamental part of crowdsourcing [175]. I designed my system to manage PFTs as a form of expert crowd where individual workers cannot be relied upon to maintain project-based knowledge and new workers must be trained quickly to perform the required labor.

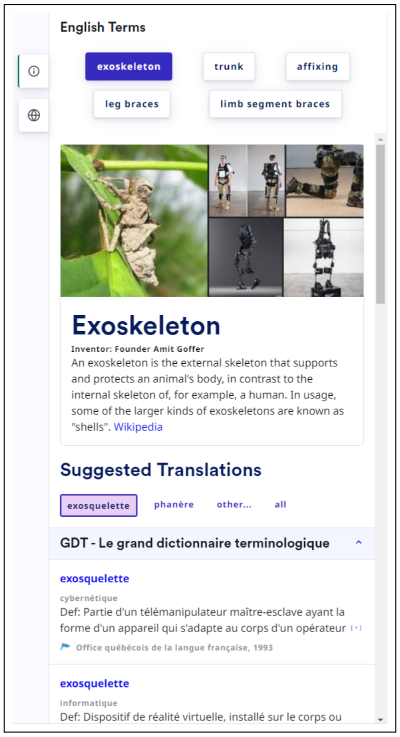
In the next section, I will discuss the details of my system and how it accomplishes these goals.

5.4 System Design

This section describes a system that meets the design goals I established from my contextual inquiry. Specifically, my system introduces an approach I call *guided context building* that makes it possible to direct PFTs' context building effort toward specific terms, potential translations, and resources. I implemented my approach on top of Lilt's commercial CAT tool[7] that features state-of-the-art neural machine translation. I added modifications to this tool that enable it to automatically summarize relevant contexts for the NMT suggestions within a *context panel*. I will first discuss the relevant features of the existing CAT tool, and then I will explain the modifications I made.

My CAT tool uses an interactive translation workflow (Figure 5.4). As a PFT translates, the system provides them with a next-word suggestion that they can choose to accept (and auto-complete the word) or reject by entering their own translation. The system automatically adapts to the prefix of what a PFT has already entered, so suggestions can only build off of a PFT's existing translation.

[7] https://lilt.com/


English Terms

Exoskeleton

Inventor: Founder Amit Goffer

An exoskeleton is the external skeleton that supports and protects an animal's body, in contrast to the internal skeleton of, for example, a human. In usage, some of the larger kinds of exoskeletons are known as "shells". Wikipedia

Suggested Translations

GDT - Le grand dictionnaire terminologique

exosquelette

cybernétique

Def: Partie d'un télémanipulateur maître-esclave ayant la forme d'un appareil qui s'adapte au corps d'un opérateur [+]

Office québécois de la langue française, 1993

exosquelette

informatique

Def: Dispositif de réalité virtuelle, installé sur le corps ou

Figure 5.5: Context panel: Based on my findings from the contextual inquiries, I developed a guided context building approach which automatically shows relevant information for terms as the PFT works on a sentence. The context panel combines information found from many resources, such as dictionaries, Wikipedia entries, concordance results, and alternative MT suggestions.

I augmented this existing CAT tool to incorporate a context panel that includes a summary of the context for several of the most difficult-to-translate terms within the text (Figure 5.5). For each term, I included a Google search-style preview of the English term and a list of potential translations compiled from several popular translation resources. Based on the findings from my prestudy, I selected four popular resources for English-to-French translation whose results I compiled in the context panel including the *Grand dictionnaire terminologique* (GDT), *Termium*, *WordReference*, and *Linguee*. I scraped these resources for their relevant information about terms and translations, and I provided them in a unified format that PFTs could quickly filter through by selecting translations from the provided list, with the MT suggestion highlighted in purple. Now that I have described the components of my system, I will now transition to describing how I evaluated the effectiveness of my approach.

5.5 Study

In this section, I will describe the steps I took to evaluate how my system meets my design goals. I did this by designing my study in two parts: first, I sought to understand the impact that my approach could theoretically have on the distribution of context building effort, and second, I sought to understand the impact of context building effort on translation outcomes such as task time, confidence, and translation quality. Specifically, my research questions are as follows:

RQ1. *What is the potential impact of guided context building on participants' context building behaviors?*

RQ2. *How do participants' context building behaviors impact translation outcomes (i.e., time, confidence, and quality)?*

RQ3. *How do participants perceive the impact of guided context building?*

My research questions are designed to evaluate all three of my design goals. Specifically, RQ1 and RQ2 are intended to evaluate different aspects of G1 (provide support for evaluating AI assistance) where RQ1 evaluates participant's usage of my tool for evaluating AI suggestions and RQ2 evaluates the effectiveness of that usage. Likewise, RQ3 is intended to evaluate both G2 (supply credibility signals from known trustworthy resources) and G3 (accommodate the conditions of crowd-based labor) by determining whether participants perceive context information to be coming from credible resources and by determining how my tool impacted participants' perceived ability to familiarize themselves with unfamiliar terms.

To answer these questions, I recruited professional English-to-French freelance translators to participate in an online controlled trial study. I will begin by discussing the nature of my study including my intended participants and general procedure, then I will cover the conditions and measures I used to answer the research questions, and I will conclude by describing the statistical methods I used to analyze my data.

5.5.1 Participants and Procedure

I recruited professional English-to-French freelance translators from UpWork [98] to participate in an hour-long online study, compensating them $30 for their time. I selected participants that had earned at least $100 on the UpWork platform doing translation tasks and had a job success rate of at least 70%. I used the same three text snippets as in my prestudy for the experimental task data.

Based on the criteria I previously described, I selected a total of 98 professional freelance translators to participate in my study. These participants had substantial qualifications that justify their selection in the study: about 60% with two or more years of experience and about 30% with more than 5 years of experience. Only 10% of my participants had less than six months of experience doing professional freelance translation. Additionally, about half of my participants had received formal education for translation; the other half were self-taught or learned through professional experience.

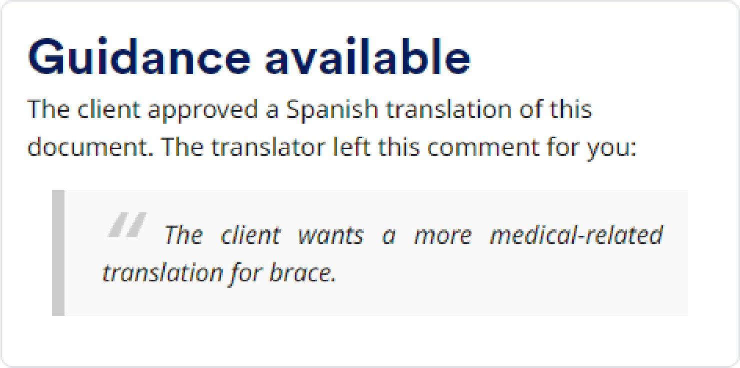
My participants had experience doing a wide variety of translation tasks including of the type I examined in my study. These included doing translation for e-commerce owners, lawyers, and medical content creators, as well as various other types of clients like search engine optimization agencies, book publishers, bloggers, and teachers. About half of my clients indicated that they specialized in a particular type of content; although, their indicated specializations were typically broad enough to include general categories such as healthcare and marketing. Additionally, while all of my participants had experience translating English-to-French, most of my participants had experience translating additional language pairs such as English-to-German and English-to-Spanish. In the next parts, I examined the effect of my approach on the context building behaviors and translation outcomes of these participants.

In the study, I asked each participant to complete a prestudy survey, a tutorial of my system, to translate four text snippets (a gold standard and the three text snippets as seen in Figures 5.1, 5.2, and 5.3), and a post-study survey. Between each translation task, I additionally asked participants to complete a short post-task survey. For each of the translation tasks, I asked participants to translate as fast and accurately as possible, using the first snippet as a gold standard (see Appendix B.1 for details). Prior to the start of the study, participants were informed about the study's requirements, purpose, procedures, and compensation, and they all indicated their consent to participate.

5.5.2 Conditions

To understand the potential impact of my guided context building approach on context building effort, I experimented with two types of guidance inspired by the findings from my contextual inquiry: 1) a *context panel* that provides relevant contexts organized by translation and 2) *targeted guidance* that provides guidance from another PFT whose translation into another language was "approved by the client." I compared these experimental conditions against two baselines that vary the amount of context participants are provided for MT suggestions. For each text snippet, I selected the 4-5 terms participants spent the most time building context for in my prestudy to include in the context panel. My conditions are as follows:

- Suggestions only (Baseline): I provided participants only with MT suggestions, no additional context, and no way to search for additional context. Participants were given an empty context panel that shows terms with potential translations, but no other information.



Guidance available

The client approved a Spanish translation of this document. The translator left this comment for you:

> *The client wants a more medical-related translation for brace.*

Figure 5.6: Guidance: For the most difficult terms, I provided guidance from another experienced PFT to help PFTs curate their context building efforts.

- Browser only: I provided participants with a built-in browser that they could use to search for additional context. I tracked both their search terms and the websites they visited while using my browser[8].
- Browser and context panel: I provided participants with a built-in browser and my context panel. For each snippet, I compiled the results of four of the most commonly used resources from 4-5 of the most commonly searched for terms in the prestudy. I additionally included a Google search style preview of the English version of each term (see Figure 5.5).
- Browser and context panel with guidance: I again provided participants with a built-in browser and my context panel, but I additionally included guidance from another PFT whose translation in another language was "approved by the client" (see Figure 5.6). For the most difficult term in each snippet, I provided guidance aimed to help PFTs curate their context building efforts toward types of translations the client is looking for[9].

I designed my experiment to tightly control the ways in which participants are able to engage in context building so that I can measure the impact of my approach on their behavior. Specifically, I asked them not to use resources outside of my web page, giving them periodic reminders of this request including at the moment they click out of my web page. Furthermore, I took the additional step of tracking cases where participants do leave the web page and monitor their behaviors when they return to ensure no external resources influenced their decision making. In conditions where

[8] Participants were repeatedly warned that I would be tracking the websites they visited and the text they entered while using my built-in browser.

[9] For each snippet, I based the exact phrasing of the guidance off of a quote from a participant in my prestudy. I kept the focus of the feedback on the client's expectations for the translated English connotations of each term so that it may feasibly be used to determine the impact that communication across translation teams may have on PFTs translating text in another language.

the participant is allowed to use the web browser, I provided them with a built-in browser that allowed them to search Google freely without leaving my web page. I assigned participants one unique condition per snippet and counter-balanced the order in which participants received conditions and text snippets. For each participant, I randomly selected from all possible orders that they could receive text snippets (6 possible orders) and from a Latin square of possible ways that they could receive conditions (4 possible orders), resulting in a repeated measures study design with 24 unique snippet-condition order combinations.

5.5.3 Measures

I measured a number of factors intended to capture multiple aspects of participant's context building effort distribution and task outcomes. My measures were as follows:

- Task time: I measured the amount of time it takes a participant to complete a translation task starting from the point that they selected "start task" in an opening dialogue menu to the point that they selected "submit"[10]. This measure does not include the time they spent completing any pre or post-task surveys.
- Browser time: I measured the amount of time a participant spent in the built-in browser by adding up all of the times from the point at which they make a search, to the point at which they continue typing in their translation. I additionally hand-coded each search a participant made within the browser for the specific resource they visited and term/translation they were looking for based on the URL they visited and the search terms they entered.
- Sidebar time: I measured the amount of time a participant spent in the sidebar by adding up all of the times from the point at which they interacted with the panel through a button click, to the point at which they continued typing in their translation. In the two baseline conditions without the context panel, the sidebar contained MT suggestions with no additional context.
- Context building time: I combined browser time with sidebar time for an aggregate measure.
- Term confidence: in post-task surveys, I asked participants to indicate how confident that their translation for each term was the "the best possible translation" for the term on a 5-point Likert scale with 1 as the least confident and 5 as the most.
- Term client-confidence: for the most challenging term in each text snippet, I asked participants to indicate in a post-task survey how confident they were that they "understand what

[10] I closely monitored for participants that experienced connectivity issues or other technical issues and removed affected data points from my analysis. Additionally, participants were not able to revisit tasks once they had been submitted, ensuring that future translation tasks had no effect on prior translation work.

type of translation" for the term the client is looking for on a 5-point Likert scale with 1 as the least confident and 5 as the most.

- Term quality: I hand-coded the translations for the most challenging term in each text snippet based on how well the translation conveys the hypothetical client's desired meaning of the English source term. I based the criteria for each term off of descriptions provided by the most experienced PFTs I interviewed in my prestudy. I coded quality for each term on a 3-point Likert scale where 0 indicates the lowest quality and 2 the highest.

5.5.4 Statistical Analysis

To analyze the impact of my approach, I constructed four linear mixed models (LMMs) predicting context building time, task time, term confidence, and term quality. LMMs are useful because they support hierarchical effects [72] and they can control for the effects of repeated measures and other variables (such as the text snippet), making them a common choice in similar studies [13]. I followed the recommended method for obtaining p-values [161] which involves comparing the full model to a version of the model without the effect in question using a likelihood ratio test. In all of my models, I controlled for the participant and text snippet to factor out confounding effects.

In addition to my quantitative analysis, I answered RQ3 by conducting a qualitative analysis of feedback participants provided me in a post-study survey. I used a similar approach that I used in my prestudy to identify patterns within comments participants left and their answers to the question "Tell us about how the CAT tool helped you translate difficult-to-translate terminology".

I designed my study to evaluate how well my approach of guided context building supports my three design goals. In the next section, I will share the main findings from my exploration.

5.6 Results

Guided context building does impact both context building behaviors and translation outcomes. This section discusses these findings and Appendix B.2 reports the statistical analyses in further detail. I will walk through my findings by answering my three research questions about: 1) how my approach impacts the way PFTs engage in context building, 2) how context building behaviors impact translation outcomes such as task time, confidence, and quality, and 3) how participants perceive the impact of guided context building.

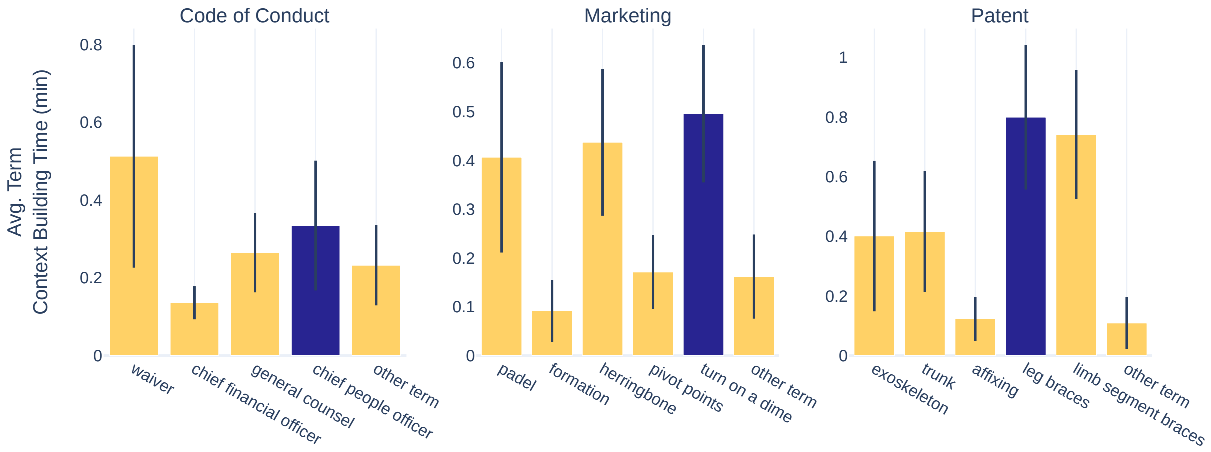


Figure 5.7: Participants spent the vast majority of their context building time on the terms I identified in the context panel. Time spent on non-context panel terms are grouped together under ‘other terms’. The blue bar shows the term I provided guidance for and 95% confidence intervals are shown by the black vertical bars.

5.6.1 RQ1: Context Building Behaviors

My approach significantly impacted participants’ context building behavior by shaping both their total context building time and the resources they used to do context building. Context building behaviors made up a significant portion of task time and they were directed towards complex terms, confirming the results of my prestudy. Second, my approach can be used to both increase overall context building time and narrow the focus of it. Finally, my approach pushed participants to use local context over external context. In addition to answering my first research question, my findings demonstrate support for my first design goal (provide support for evaluating AI assistance) based on the fact that participants opted to use my system over available alternatives (browser). I will discuss each finding in detail.

5.6.1.1 Context building time made up a prominent percentage of task time and was directed toward complex terms.

In accordance with the findings from my prestudy, context building behaviors made up a significant portion of participants’ task time and their effort was directed primarily toward complex terms. On average, participants spent 18.3% of their total task time undergoing context building behaviors including 10.6% in my context panel and 7.7% browsing the web. Following the coding process I previously described, I hand-coded each browser search a participant made and was able to identify

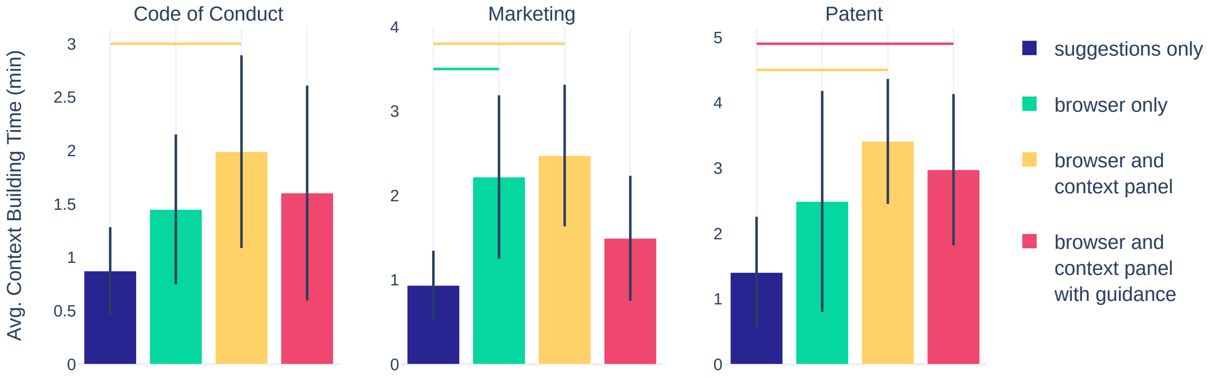


Figure 5.8: Providing context for terms (yellow bar) led to increases in context building time on all text snippets. However, providing guidance (red bar) narrowed the focus of context building time such that only the patent text snippet saw an increase. Providing participants with a browser led to an increase in the marketing text snippet but not the others. Significant statistical tests are shown by the horizontal bars at the top (colored by the significant condition on the right) as compared against the baseline (blue bar). 95% confidence intervals are shown by the black vertical bars.

the term, translation, and resource a participant used to build context for 98% of searches. Of the 2% remaining searches, 1% of them were spent searching for resources and general information about the client, and the final 1% of searches were unrelated to the translation task.

More specifically, participants spent most of their context building time building context for the terms in the context panel (see Figure 5.7 for a breakdown). Per snippet, the terms “waiver,” “general counsel,” and “chief people officer” were the most searched for in the code of conduct snippet, “padel,” “herringbone,” and “turn on a dime” in the marketing snippet, and “leg braces” and “limb segment braces” in the patent snippet. When participants searched for terms not included in the context panel, they generally spread their time across many different terms. The combination of these times is comparable to the time they spent on any one of the context panel terms.

5.6.1.2 Providing more context led to an increase in context building time.

Providing context for terms in the context panel pushed participants to spend more overall time doing context building. As is visualized in Figure 5.8, the condition where I provided participants with both a browser and context panel (yellow bar) led participants to spend more time building context on all three text snippets compared to when they have suggestions alone (blue bar). This difference in time is starkest in the patent snippet which showed a 2 minute increase in average context building time.

Digging further into where the increase in context building time occurs, the context panel did not replace the browser, but rather, it inspired more use of the browser to find supplemental context (i.e., information not found in the context panel) in addition to searching through the context panel (see Figure 5.9, right chart). In particular, the increase in context building time was most prominent when comparing the upper quartiles of each condition, where I observed more than a 10x increase in the time participants spent using the browser to search for non-context panel information in the condition where they have the context panel compared to where they have only the browser.

5.6.1.3 Providing guidance led to a narrowed focus of context building time.

On the contrary, providing participants with written guidance (a short statement about what kind of translation for a term the client is looking for) led participants to narrow the focus of their context building by spending less time building context for some terms. As is seen in Figure 5.8, participants in the condition where I provided them with access to a browser, context panel, and guidance (red bar) spent more time building context for terms in the patent text snippet, but a statistically insignificant difference in time for the code of conduct and marketing text snippets as compared to the condition where they have only suggestions (blue bar). Additionally, guidance often led participants to exert context building effort in different places than in the condition where I provided them with a browser alone. Specifically, participants in the browser condition (green bar) spent more time building context in the marketing snippet compared to the baseline where the guidance condition did not see an increase, and participants in the guidance condition spent more time building context in the patent snippet compared to the baseline where the browser condition did not see an increase.

Digging into where the narrowed focus in context building occurred, guidance reduced the amount of time participants spent searching for information not provided in the context panel (i.e., searching for information about translations I suggested using resources I provided), as can be seen in Figure 5.9 (right chart). This decrease in time was most noticeable within the upper quartile of participants who spent nearly one quarter of the time in the browser searching for non-context panel information in the guidance condition compared to the context panel condition.

5.6.1.4 Participants used local context over external context.

Looking closer at the resources participants used to build context, participants generally favored searching through local context (context provided in the context panel) over external context (context found from searching in the browser). As can be seen in Figure 5.9 (left chart), participants

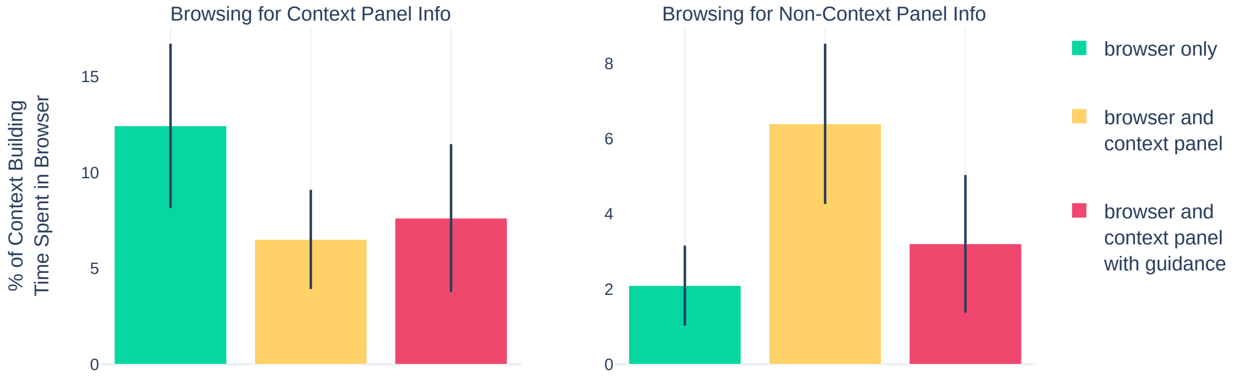


Figure 5.9: I show percentage allocation of context building time spent in the browser to search for information they could find in the context panel (left chart) and information they could not find in the context panel (right chart). I found that participants used local context (context provided in the context panel) over external context (context found from searching in the browser) and that increases in context building time in the condition where they had access to the context panel mostly manifested through an increase in browser searches for non-context panel information in the upper quartile of participants. 95% confidence intervals are shown by the black vertical bars.

in conditions where I provided them with the context panel decreased their use of the browser to search for information that could be found in the context panel (i.e., search for information about translations I suggested using resources I provided).

This finding demonstrates that participants opted to use my tool for evaluating AI suggestions in place of the standard available alternative (browser). In particular, when participants sought to gather context surrounding terms or translations, they traded the time they would otherwise spend searching for context using the browser by instead searching through the context I provided them in the context panel. This traded time is evidence that my system provides support for evaluating AI suggestions (meeting my first design goal) since participants were able to find relevant context for AI suggested translations in the context panel, eliminating their need to find it elsewhere. Now that I have described the impact of guided context building on participants' context building behaviors, I will next explain how context building behaviors impact translation outcomes.

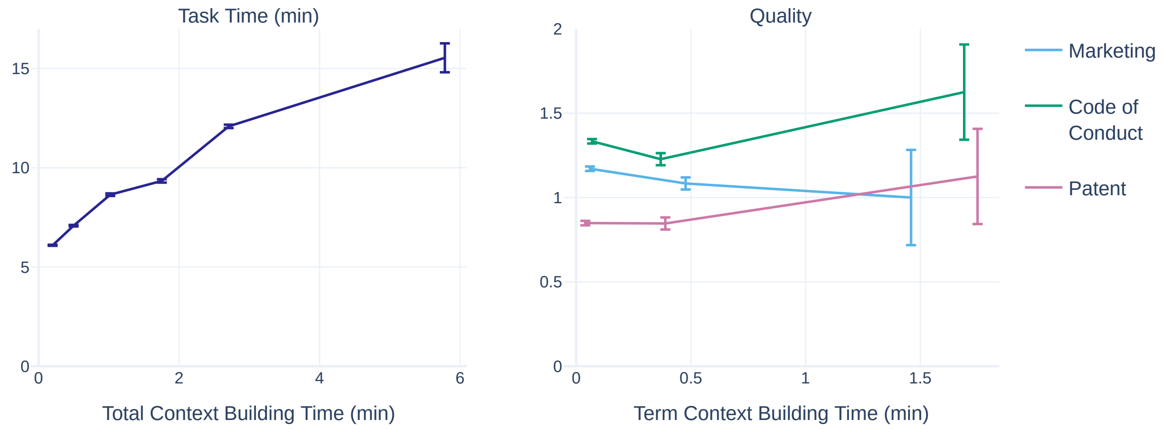


Figure 5.10: The translation outcomes task time (left chart) and quality (right chart) are predicted by context building time. For task time, there is a strong positive correlation where participants who spent more time building context took longer to complete the task. For quality, there is a slight positive statistically significant correlation for the code of conduct and patent text snippets, but not the marketing text snippet. 95% confidence intervals are shown by the vertical bars.

5.6.2 RQ2: Translation Outcomes

I found two ways in which context building effort impacts translation outcomes including by increasing task time and by improving quality for some terms (see Figure 5.10); however, I found no effect on confidence. For each outcome, I computed a separate linear mixed model predicting the relationship between context building time and the outcome in question. My findings demonstrate that I have met my first design goal in some cases; I will explain in more detail. I report each of these models in Appendix B.2.

5.6.2.1 Task time is tied to context building time.

My first outcome, task time, is strongly correlated with the time participants spent building context (Figure 5.10 left chart). This relationship was very strong: participants that spent the most time building context took nearly 3 times as long to complete the task as participants that spent the least amount of time building context.

5.6.2.2 Context building led to better quality translations for some terms.

For some terms, quality is also correlated with the time participants spent building context for the term. Specifically, when I examined the relationship between the amount of time participants spent building context for each of my pre-selected guidance terms (i.e., the terms I provided guidance for) and the quality of translation they select, I found a significant positive interaction effect for the code of conduct and patent text snippets, and no significance for the marketing snippet. As is shown in Figure 5.10 (right chart), participants who spent a longer time building context for the terms 'chief people officer' (code of conduct) and 'leg braces' (patent) were more likely to score higher in quality for those terms. I note that this effect was small, the R^2 for the model is only 0.09 indicating that context building time only slightly predicts quality.

This finding shows partial support for my first design goal (providing support for evaluating AI assistance). Specifically, this finding demonstrates that by using a guided context building approach, participants' context building effort can be tailored to spend additional context building time where it needed. That additional time is then more likely to lead to improved quality for some terms such those present within the patent text snippet.

5.6.2.3 Confidence is independent of context building time.

Despite it taking participants longer and being more likely to see improvements to quality, the time participants spent building context was not predictive of their confidence. More specifically, I saw no statistically significant relationship between term context building time in relation to participant's confidence that their translation for the term "is the best possible translation" or that they "understand what type of translation [for the term] the client is looking for". My model has substantial power: using a simulation-based power analysis[11] as is recommended for linear mixed models [30] with a total of 1,342 unique confidence ratings across 98 participants and 14 terms[12], I estimate that I can detect a relationship as small as a 0.1 increase in confidence scores (on a 5-point Likert scale) for every minute increase in context building time with 90.8% power. My power increases to $> 99\%$ for slopes greater than 0.15.

As I have seen in my analysis, context building significantly impacts some translation outcomes (i.e., task time), has a complex relationship with others (i.e., quality), and has very little or

[11] Coefficients for simulation-based power analyses are typically selected from computed models in a pilot study or from prior literature [30]. However, with a lack of sufficient pilot data and relevant prior literature, I instead used coefficients from a model computed on my dataset. To combat potential bias that could result from this approach, I performed a sensitivity analysis with these coefficients and found that small differences in the coefficient values generally did not impact the result.

[12] While a complete dataset would be comprised of 1,372 unique ratings, as I have previously explained, a few of my participants experienced connectivity issues and other data problems on some tasks that forced me to drop those tasks from the analysis.

no relationship with others (i.e., confidence). I will next look into the factors that explain these relationships as can be seen from participant's qualitative responses.

5.6.3 RQ3: Perceived impact of guided context building

Participants left comments that explained how my approach shaped the context building process. I gathered qualitative feedback at the end of the study by asking participants to "Tell us about how the CAT tool helped you translate difficult-to-translate terminology." From this data I found that my approach shaped participant's understanding of terms and potential translations, which was most helpful in kick-starting the context building process in what would otherwise be unfamiliar domains of translation. However, many still distrusted the suggestions and thereby used the provided context as a launching point for further investigation. My findings show that I have met my second (supply credibility signals from known trustworthy resources) and third (accommodate the conditions of crowd-based labor) design goals. I will discuss these more in detail.

5.6.3.1 Providing context shaped participant's understanding of terms and translations.

Participants indicated that my provided context for both the English term (i.e., English definitions and guidance) and translation suggestions influenced their understanding of terms. In cases where the English term is unfamiliar, participants mentioned that they used the translation suggestion list to gather context clues about its meaning, and they used the number of examples for suggested translations to decide where to focus their effort. Additionally, some mentioned that they used the provided guidance to clarify what was expected for translations of specific terms.

Furthermore, participants thought that the provided sources for context around translation suggestions was particularly helpful for selecting a translation. Specifically, many mentioned that the diversity of resources represented (and resulting diversity of translation suggestions) made it possible for them to find translations they may not have otherwise considered on their own. Likewise, participants expressed that sourcing the context helped them develop a deep understanding of translation suggestions:

> It is also very important for me to mention that having worked with similar CAT tools in the web, none of them has the VERY useful feature of accessing the source material as this one does. It allows [me to understand the] real context of [translation suggestions] helping this way to build up real, deep meaning for the potential target reader.

5.6.3.2 Providing context helped familiarize participants with technical terms in unfamiliar domains of translation.

I observed the most impact for participants completing tasks in domains they were not familiar with, particularly when they translated technical terms. Many participants noted that the provided context helped them assess the credibility of translation suggestions, which was useful for technical terms that have precise meanings:

> For technical terms, especially those related to the translation of the Code of Conduct, the examples given by the CAT tool were very useful in choosing the correct translation with professional qualifications that don't have a direct translation in French (i.e General Counsel). [...] Example sentences are very useful for technical terms, especially medical ones (exoskeleton translation), to know how these words are used and what their exact translation is in the medical jargon for example. For a translator new to this field, it can be complicated to translate these terms, so it is a really valuable tool.

5.6.3.3 Many participants showed signs of distrust in the provided translation suggestions, even if translation outcomes ultimately improved.

Despite numerous comments that my context and my suggestions were helpful, many participants still indicated that they both distrusted my tool and that they went externally to verify potential translations. In particular, I saw evidence that participants distrusted two aspects of my tool: 1) the quality of the suggestions themselves and 2) the completeness of the suggestion list. Many participants indicated that they thought my suggestions gave a few ideas of some possible translations, but they would then continue investigating from there. Others indicated a desire to see a greater depth of context for the suggestions, such as discussions from other PFTs:

> The suggestions made by the CAT were inspiring, sometimes 100% accurate. But sometimes, some terms were impossible to translate (as always), and I do enjoy browsing forums such as Wordreference where translators discuss expressions. It either gave solutions or helped in thinking outside the box and the accurate translation.

An interesting outcome from this distrust is the seeming disconnect between how translators perceived the context panel's information and their increased performance on the quality measures in the guided condition. I speculate on this in the next section.

I have demonstrated that participants altered their approach to context building as a result of my tool. Much of this change in behavior resulted in an improved understanding of proposed translation suggestions — particularly when translating technical terms in unfamiliar domains of translation — but it did not satisfy all of participant's needs. In the next section, I will recap how my findings tie back to my contributions and conclude.

5.7 Discussion

In this work, I have examined the role AI could play in supporting context building for collaborative human-AI social reference processing systems. I began by conducting a prestudy with PFTs where I examined how they engage in context building with the assistance of MT. From the interviews, I showed that context building is essential for evaluating the utility of MT suggestions and for finding high quality translations for complex terms. I additionally formulated three design goals for a proposed system including G1 (provide support for evaluating AI assistance), G2 (supply credibility signals from known trustworthy resources) and G3 (accommodate the conditions of crowd-based labor).

I then designed an approach to support PFTs by meeting my design goals that I call *guided context building* which directs context building effort toward examining specific terms, connotations, and resources. I implemented this approach on top of Lilt's existing commercial CAT tool and performed a user study to evaluate its efficacy for shaping PFTs' context building effort.

To evaluate whether my guided context building approach meets my design goals, I asked three research questions: RQ1) What is the impact of the guided context building on context building behaviors? RQ2) What is the impact of guided context building on translation outcomes? and RQ3) How do participants perceive the impact of guided context building? Through my experiment, I showed that it is feasible to shape both the amount and type of context building PFTs engage in. By providing PFTs with access to a context panel featuring findings from commonly used resources, I pushed PFTs to do more context building using the context I provided them (RQ1). Likewise, by providing PFTs with targeted guidance that offers insights into what the end-client wants for a translation, I pushed PFTs to narrow the focus of their context building — leading to a reduction in context building effort for some terms (RQ1). In both cases, I demonstrated support for G1 in that PFTs opted to use our tool over an available alternative. Additionally, I demonstrated that context building impacts downstream translation outcomes by increasing overall translation time and translation quality for some terms (RQ2) — partial support for G1 in the cases where performance increased. Finally, I showed support for G2 and G3 by exploring how PFTs perceive the impact of guided context building (RQ3), demonstrating that PFTs perceive my context to be credible due to being backed by known trustworthy resources (G2) and that PFTs found my approach to be particularly beneficial for on-boarding in new domains of translation (G3).

Based on the findings from my prestudy and evaluation of guided context building, I argue for the importance of building collaborative tooling that assists people in determining where they should exert their effort in evaluating AI suggestions. I have supported this argument with evidence from my prestudy that showed how evaluating AI assistance became the focus of PFT effort so that they can maintain consistency with other PFTs. Additionally, I demonstrated how guided context building can be implemented to shape evaluative effort to achieve the goals of a system designer.

My contributions are therefore: 1) a demonstration of the importance of building tooling that supports the human effort required to evaluate AI-suggested context within collaborative human-AI social reference processing systems, 2) an exploration of how PFTs engage in process of context building when working in collaboration with MT, 3) an guided context building approach for shaping human context building effort, and 4) a system that incorporates my idea of guided context building for language translation and an evaluation of my approach with PFTs where I demonstrated its feasibility to shape context building effort.

My work demonstrates a case study of how human effort within collaborative human-AI social reference processing systems can be shaped to improve task outcomes such as speed and quality. In my domain of language translation, I witnessed multiple examples of PFTs exerting a great deal of effort to understand the MT suggestions even when it did not lead to a measurable improvement in the quality of translation they selected. I argue that guided context building could be used to re-calibrate human evaluative effort in these cases and in other cases where external search is necessary, as I have found that it can shape a variety of context building behaviors.

I additionally argue that guided context building can be implemented to shape context building effort according to one's desired goals for the human-AI social reference processing system. I experimented with two forms of guided context building: one that automatically curates context from popular resources and one that provides shorter, more targeted guidance. I found that the first form (contextual information without targeted guidance) can be used to increase context building activities and that the second (with targeted guidance) can be used to focus effort in specific places. While I only simulated targeted guidance in this chapter, I open numerous opportunities to explore how such guidance can be constructed and shared across human-AI teams through knowledge sharing, as has been done in other CSCW domains [1].

My findings indicate a need for further research into the factors that influence human effort within human-AI collaborations. Guided context building can be thought of as a form of explanation for AI suggestions such as those studied by Bansal et al. [17] and Buçinca et al. [31]. However, I add nuance to the findings of these prior studies in my finding that providing more contextual information led to an increase in overall context building effort — and therefore *reduced* their reliance on the AI — rather than resulting in overreliance. I argue that this difference in reliance behavior can be explained by the PFTs' aversion to taking risks, as their job livelihoods (e.g., payments for the translation job) are dependent on the quality of their output translations.

My work presents opportunities for AI to make improvements to people's overall performance on a task, even if they cannot perceive it. In my study, I have seen evidence of a disconnect between how translators perceived the accuracy of the MT suggestions (as displayed in their distrust), their perception of their selected translation quality (as measured by confidence), and their actual translation quality. Even though translators might have distrusted the MT suggestions and perceived their confidence to be low, their actual translation quality still improved in some cases. This result builds on the finding of Glikson and Woolley [77] in that people's perceptions about the AI system (trust) can predict their level of reliance on the AI system, but reliance may be more tightly linked to system outcomes. I argue that my quantitative measure for inferring human reliance on the AI system (tracking context building time) can be more accurately used to predict system outcomes.

While I have examined one domain in this chapter, language translation, my findings apply to any context building process conducted by a social reference processing system. For example, in addition to assisting with detecting emotionally manipulative language and othering rhetoric, my approach might also be applicable in cases where AI assists with creative writing — where it may generate new social references from mash-ups of old ones. I believe that my approach will be important for designing AI systems that can assist humans in delineating these connotations, leading to more reliable human-AI collaborations.

5.8 Conclusion

In this chapter, I have contributed a prototype tool that provides AI support for gathering the connotations surrounding social references. My approach, which I call *guided context building*, shapes human effort within collaborative human-AI social reference processing systems by directing context building effort to places where it is needed the most. Using findings from a series of semi-structured contextual inquiry interviews with professional freelance translators (PFTs), I proposed and implemented an approach that directs PFT's toward examining specific terms, translations, and resources. I evaluated my approach through a user study of 98 PFTs in three domains of translation and showed its feasibility for shaping context building effort. I demonstrated that PFTs engage in context building in a complexity of ways, including to understand the utility of MT suggestions. I argue that guided context building can be used to improve human-AI collaboration outcomes (e.g., speed and quality) by focusing effort in places where it is needed most. More generally, I argue that it is important to develop collaborative tooling that assists people in determining where to exert effort toward evaluating AI-suggested context. Additionally, I argue that guided context building is a useful approach for providing AI support for building context around social references that complements alternative approaches such as recruiting expert panels.

CHAPTER 6

Conclusion

In this dissertation, I have introduced the concept of a social reference and I have worked toward building collaborative human-AI systems that can do effective social reference processing. In this chapter, I will begin by summarizing my main findings, then I will reflect on how my findings can be put together into an effective social reference processing system, then I will discuss potential limitations of my approach, and I will conclude by suggesting directions for future research.

6.1 Summary of Findings

In this section, I will summarize the arguments I have made thus far by first revisiting the framework I use to understand social references and social reference processing, and then I will revisit my research questions and summarize my main contributions.

6.1.1 Revisiting a framework for social reference processing

I have based my idea of a social reference and social reference processing on concepts from both semiotics and its sub-domain, conceptual metaphor theory. Specifically, I used Kövecses's catalog of contextual factors that influence metaphorical conceptualization [120] to understand the kinds of context necessary to understand social references and Odgen and Richard's idea of the semiotic triangle [169] to understand the subprocesses necessary to process them. Using this basis, I created a useful framework to conceptualize the social reference processing research space and to develop the research questions I explored throughout this dissertation.

6.1.1.1 The social reference processing triangle

In chapter 1, I first proposed that social references are a type of symbol that can be understood through the process described in the semiotic triangle. As Odgen and Richards conclude in their work [169], symbols do not carry inherent meaning, but instead are understood through a process

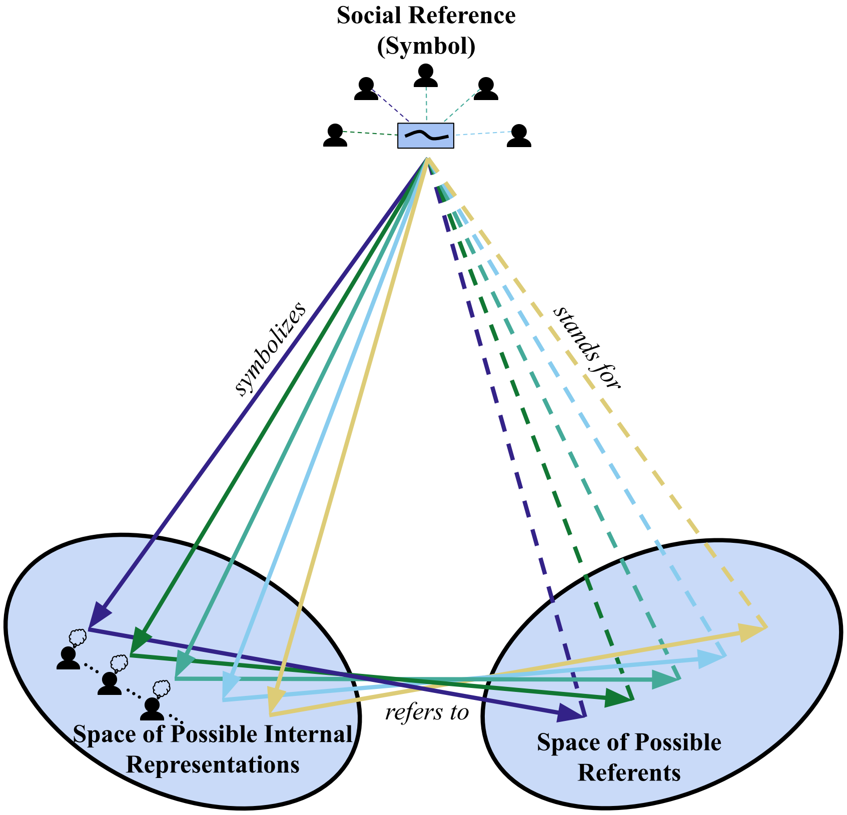

Social Reference
(Symbol)
symbolizes
stands for
refers to
Space of Possible Internal Representations
Space of Possible Referents

Figure 6.1: Social reference processing builds on Ogden and Richards' idea of the semiotic triangle [169]. Individuals interpret social references (symbols) into a space of possible internal representations. Each of those internal representations refers to a referent (connotation or group of connotations) in the space of possible referents. In short, social reference processing involves using groups of people to gain an understanding of the space of possible connotations a social reference could represent.

that involves encoding them into an internal representation (by people) which refers to a concept (referent). This interpretation process is developed through socialization, which allows people to associate symbols (received orally, visually, or through any system of signals) with meanings.

Social references are unique in that they may be encoded into numerous possible internal representations that differ between people, and thereby, they may refer to many possible connotations. I illustrate the significance of this feature in Figure 6.1 where I argue that developing a collective understanding of social references involves developing a *space of possible internal representations* and a *space of possible referents*. Social reference processing involves building an understanding of these possibilities by gathering the perspectives of many agents (people) to interpret the social reference, either directly through recruitment or indirectly through the use of tooling.

6.1.1.2 The subprocesses involved in social reference processing

In chapter 1 I additionally used this basis for understanding social references to argue that social reference processing involves at least three subprocesses: extraction, interpretation, and action (see Figure 1.1). The first subprocess (extraction) is necessary for any social reference processing system that seeks to detect or evaluate the impact of social references apart from other aspects of content. For example, as I demonstrated in chapters 3 and 4, detecting emotionally manipulative language and othering rhetoric involves disentangling the impact of social references from intrinsically emotional content and other aspects of language. The second subprocess (interpretation) is another key aspect of social reference processing tasks that involves building an understanding around the space of possible connotations social references can refer to. Additionally, as I suggested in chapter 4, interpretation may also involve building an understanding around the kinds of skills, knowledge, and experiences people need to understand specific social references. Finally, the action subprocess is needed for any social reference processing system that aims to use insights about social references in a decision-making process, such as crafting interventions to prevent the harms of content containing emotionally manipulative language or othering rhetoric.

6.1.2 Research Questions

Using this framework for understanding the social reference processing space, I developed a main research question to guide my exploration:

> *How can we effectively and reliably process social references?*

I additionally used my framework to develop three additional research sub-questions that each explore a different aspect of social reference processing. For each sub-question, I selected a social reference processing problem domain where I could thoroughly explore the question and gain insights that may apply to any social reference processing task. I will briefly summarize my key findings in each domain and explain how they answer my research questions.

6.1.2.1 Disentangling social references from content

> **RQ1:** How can we disentangle the effect of social references from other factors within content by mitigating the biasing effect those content-factors might have on human annotators?

My first research sub-question (RQ1) was intended to explore a primary challenge any social reference processing system operating on real-world content would need to overcome: disentangling social references from content. I explored this challenge in chapter 3 by examining how to distinguish a kind of manipulative social reference, *emotionally manipulative language* (EML), from a related attribute in content that I call intrinsically emotional content (IEC). I demonstrated that human annotators are inclined to conflate EML with IEC, which can lead to false positive classification of emotional, but not manipulative, content. I introduced an approach that can make it possible to overcome this conflation error that I call *anchor comparison*. My approach involves coordinating human annotators to create an anchor version of content, where the target social reference is removed, to use as a point of comparison against the original content. I demonstrated that this approach is effective for mitigating conflation error by introducing and evaluating a system that embodies the approach.

While I argued that this approach is effective for detecting the majority of cases of EML, I acknowledged that it is limited by the knowledge and skills of the crowd workers that are recruited to operate the system. Specifically, EML elicited through subtle social references may be less reliably detected due to the possibility that the people operating the system do not have the relevant context available to understand potential connotations. I attempted to alleviate this limitation through my studies in chapters 4 and 5 where I explored two directions for accumulating relevant context.

6.1.2.2 Exploring social reference processing expertise

> **RQ2:** What expertise is necessary to combine and reason with the context surrounding social references to understand their situated meaning, and how can we identify people with such expertise from a generalized crowd?

In chapter 4, I sought to answer RQ2 by developing an understanding of the aspects associated with social reference processing expertise by exploring the challenges associated with building expert panels that I call *Justice Panels*. In my study, I examined the cues and processes experts in social justice use to find othering rhetoric. My analysis uncovered three specific insights that system designers should consider when building such panels, including that 1) relevant expertise is multifaceted, contextual, and situated, 2) that sources of disagreement among workers can come from at least two sources including *expertise divergence* and *impact disagreements*, and 3) that anchor comparison can also be used to maximize each panelist's performance. In addition to these findings, I argued that the dataset I created could be used as a gold standard for finding people with substantial relevant lived experiences from a generalized crowd. More generally, I suggested that it may be possible to use the method I demonstrated throughout my study to develop gold standards for a range of social reference processing problems, which would allow for the development of systems that can engage more deeply with panelists lived experiences.

However, as I also discovered in chapter 4, identifying crowd workers with relevant expertise is not a trivial task. While my work takes steps in this direction, I find it unlikely that a system would be capable of consistently creating panels with top experts. However, my approach might be capable of selecting panelists with adequate expertise, and it may be possible to boost their expertise by leveraging AI assistance. For this reason, in chapter 5 I explored the possibility of using AI assistance to gather potentially relevant context for social references that could be suggested to them.

6.1.2.3 Supporting the context building process with AI tooling

> **RQ3:** What role can AI tooling play in ensuring that adequate context surrounding social references is available for human annotators to parse?

Finally, in chapter 5, I answered RQ3 by exploring how to support an approach for gathering the relevant context necessary to understand social references without the need for human annotators to know it ahead of time. My approach, which I call *guided context building*, enables a crowd — who might otherwise not have the knowledge of specific references available — to strategically forage for relevant context using resources accumulated by an AI or by acting from the leads provided by other crowd workers. This approach would be effective for building context around phrases such as “links to billionaire George Soros” where previously documented information could be used to deduce the conspiratorial connotations. More generally, my approach could be useful for uncovering many kinds of situational and discourse context [120], which may include artifacts, events, prior discourse, and other aspects of culture that can be gathered and curated by an AI.

6.2 Social Reference Processing with Collaborative Human-AI Systems

As a result of my work, I have come to the conclusion that, by strategically recruiting, coordinating, and supporting crowdsourced panels of people with AI assistance, we can make effective collaborative human-AI systems for processing social references. I have come to this conclusion after having repeatedly demonstrated the pitfalls of alternative approaches, such as using computational approaches or crowdsourcing approaches alone.

While a naïve system designer might seek to do accomplish social reference processing tasks using computational approaches by themselves, evidence from my studies suggest that such a system would struggle with each of the three social reference processing subprocesses I have discussed. More specifically, I have made the argument that since automated approaches make inferences based on prior training data, they may struggle to make accurate inferences for social references that are not accurately captured in their training dataset. In chapter 3 I provided some evidence that this limitation may impact the extraction subprocess since existing automated approaches a system designer may use to identify EML were not able to consistently identify it. Additionally, I made the argument in chapter 4 that existing approaches may struggle to identify novel othering narrative archetypes as they may not fit the mold captured within prior training data — an significant limitation that would prevent early detection of new forms of othering.

I have also made the argument that, while relying on crowd approaches by themselves may capable of overcoming some of the limitations of automated approaches, they too would struggle with the extraction and interpretation subprocesses in cases where crowd members do not have adequate expertise to the understand social references they are employed to process. I provided evidence of this weakness in chapter 3 where I found that non-expert crowd workers were prone to conflating EML with IEC, and again in chapter 4 where I cataloged several forms of skills, knowledge, and experiences expert crowd workers used to find othering rhetoric. In particular, I provided evidence that it would be a substantial challenge to consistently recruit top experts from a generalized crowd, which would be necessary to understand the full space of relevant interpretations for many social references.

Instead, I have provided evidence that collaborative human-AI approaches are more appropriate for most social reference processing tasks since their performance can be bolstered with strategic recruitment, coordination, and AI support. I have demonstrated this claim by providing three contributions to the social reference processing space: 1) a new approach for coordinating crowd workers called *anchor comparison*, 2) a method for developing gold standards that engage deeply with panelists' lived experiences, and 3) a new approach for supporting the context building process with AI tooling called *guided context building*. With each of these three contributions, we can better support each of the three social reference processing subprocesses.

I demonstrated the potential of using anchor comparison to bolster the extraction subprocess in chapters 3 and 4 where I showed how it can help non-expert crowd workers differentiate between EML and IEC, and I showed how it can improve even expert performance by engaging a broader range of skills. I envision anchor comparison as being an important coordination tool for weaving together the work of human annotators within a collaborative human-AI system. To complement the approach, my method for developing potential gold standards that I showed in chapter 4 could be used to recruit an expert panel of people with relevant lived experiences. Through the use of expert panels, it is possible to bolster all three subprocesses since experts would have a better ability to recognize, interpret, and understand the impact of social references. Finally, as I showed in chapter 5, my guided context building approach could be used to provide AI support to an expert panel, which could improve their ability to interpret social references by identifying the most relevant connotations associated with them. I envision that a collaborative human-AI system that uses all three of these contributions would be capable of effectively and reliably accomplishing many social reference processing tasks.

6.3 Limitations

While I have begun to address many of the limitations that exist within current systems, my systems are not without their own limitations. Particularly, there are three persistent limitations I call attention to: 1) prohibitive costs, 2) complex disagreements, and 3) shifts in the meaning of language. I will discuss each in more detail.

I have taken many steps to expand the capabilities of human-AI systems in the social reference processing space in part to reduce costs while maintaining high quality; however, these costs may still be too high for some applications. Specifically, I acknowledge that any cost may be prohibitive in some contexts and thereby lead some to rely solely on AI-based solutions, which are prone to errors. With that mentioned, I argue that it may be possible to reduce costs by incorporating the interactions of existing users into the system itself, such as those introduced in [214]. My work takes the first steps toward this goal by suggesting how recruitment and coordination can be done from a generalized crowd. Additionally, as I showed in chapter 3, many social reference processing tasks may not need top expertise, and so in these scenarios it may be possible to reduce costs by recruiting fewer experts or even non-experts.

Additionally, in chapter 4, I identified cases where understanding social references can be particularly challenging due to the potential for complex disagreements among crowd workers. In my study, I documented instances where participants with shared niche expertise still disagree about the connotations behind specific language. In one example, I found that two participants disagreed about the racism and prejudice present in the phrase "His mother was addicted to crack" despite both having experiences with a parent struggling with drug abuse. In these cases, I found it unlikely that a system would be capable of wading through the complex social dynamics that lead even experts to disagree. However, I have suggested that it may be possible to provide early indicators when these cases occur by creating an indicator that makes use of the lack of structure in the crowds' annotations. It may also be possible to identify additional forms of disagreement that can impact the crowds' annotations and I believe that additional investigations are warranted.

Finally, while my systems have the potential to conduct accurate analysis at any particular point in time, their outputs could become stale if the meaning of the underlying language they analyze shifts after the fact. This problem is exacerbated in the case of purely automated systems since they may rely on outdated training data; however, I acknowledge that even human-AI systems cannot foresee future language shifts. For this reason, I believe that additional research into time-sensitive expertise could be fruitful for improving recruitment practices. In cases where systems rely on AI assistance, it may be useful to keep time and context information available in training data so that it can be later evaluated for potential expiration. Old training data could be marked as out-of-date or nullified entirely in cases where new context significantly differs from old context.

6.4 Future Directions

In addition to the research I have discussed in this dissertation, there are many interesting longer-term directions for future research in the social reference processing space. I will cover a few possible directions that have merit including: processing *semiotic payloads*, developing measures for expertise, and providing AI support for using social reference interpretations to take action.

6.4.1 Processing semiotic payloads

In chapter 4, I discovered that social references are a sub-class of a broader set of problems that I call *semiotic payloads*. While social references refer to *terms* that invoke connotations by overlapping parts of social and cultural contexts, semiotic payloads may include any kind of language that smuggles in connotations with a literary device, including narratives and stereotypes — not necessary contained within a word or phrase. Processing semiotic payloads may be a substantially more challenging task, as their highly unstructured nature makes it difficult to design tooling around them and they may require substantially more effort from people to interpret. Additionally, semiotic payloads make all three of the social reference processing subprocesses more challenging to complete as they are even more intertwined with existing content than social references and they often require more expertise to understand.

6.4.2 Developing measures for expertise

While I have introduced potential gold standards for identifying people with relevant lived experiences in chapter 4, an important area of future research will involve developing more formalized measures of expertise for social reference processing problems. In the social justice context, I have identified numerous forms of expertise that experts use to find othering rhetoric, most prominently including media literacy skills and relevant lived experiences. A more developed measure would make it possible to precisely evaluate potential workers' ability across these and additional factors, which could help identify top experts from a generalized crowd. I believe that developing these measures for a variety of social reference processing tasks will be important for ensuring that systems can efficiently and quickly recruit relevant expertise.

6.4.3 AI support of using social references to take action

In addition to the guided context building approach that I introduced in chapter 5 that bolsters performance on the interpretation subprocess, it will important to develop approaches for providing AI support to bolster performance on the action subprocess. This direction may involve using the interpretations of social references garnered through the approaches I have already introduced to craft personalized interventions or guidance on decision-making. For example, a content moderator system may use AI support to suggest potential rule violations to human moderators or suggest possible actions that can be taken to reduce the harm of the content.

6.5 Conclusion

I have introduced three approaches for managing the challenges of social reference processing. I believe social reference processing systems will be critically important for addressing numerous problems in the online world, including manipulative and othering content. For both of these domains, existing solutions rely on a find-and-delete paradigm, where content that is egregious enough to be consistently detected is removed. My hope is that social reference processing will extend the capabilities of these systems to enable applications that provide agency to users so that they can be armed with accurate information and context to turn away from such content on their own. At the same time, my approaches for building context around social references can be broadly applied in other domains such as the one I studied in chapter 5 (language translation) or potentially for building shared context in collaborative writing settings. My hope is that social reference processing will enable the use of computation to improve our collective understanding of language and create a more healthy and creative society.

APPENDIX A

Appendix of Chapter 4

A.1 Text Snippets

The following are the text snippets I used in my analysis in chapter 4:

A.1.1 Snippet 1

> Six illegal aliens have been charged with sexually assaulting and trafficking a girl under the age of 16 in Wisconsin. All of the six men arrested are believed to be illegal aliens from Guatemala, according to the Green Lake County Sheriff's Office. Police said the underage girl called 911 as she was on the run after escaping from the six illegal aliens. After the call, police seized evidence they believes reveals the extent to which the girl was being sex trafficked and sexually assaulted by the illegal aliens. Illegal alien Evis Amabilio Garcia-Rivera is suspected of sex trafficking the girl. Currently, the illegal aliens are being held on bail at the Green Lake County Correctional Facility.

A.1.2 Snippet 2

> A pro-mass immigration organization with links to billionaire George Soros has successfully lobbied six Republican governors to resettle more refugees in their states. The federally mandated refugee resettlement program has brought more than 718,000 refugees to the U.S. since January 2008 – a group larger than the entire state population of Wyoming. Refugee resettlement costs American taxpayers nearly $9 billion every five years, according to the latest research. Over the course of five years, an estimated 16% of all refugees admitted will need housing assistance paid for by taxpayers.

A.1.3 Snippet 3

Caldwell’s story is remarkable and inspiring. One of nine children raised in poverty on the South Side of Chicago, he had to face true adversity. His mother was addicted to crack. His community was poor, rundown and violent. Caldwell recalled to me how he used to think Republicans belonged to the party of racists who didn’t care about minorities or the poor. But then one day, an elderly Black man on the South Side challenged him, explaining that at one point most Black Americans were Republicans and conservatives. So, Caldwell did a little research and realized the man wasn’t lying. So began his transformation to become a Black conservative. Interestingly, Caldwell explained that many Black Americans are actually conservative and believe in family values – they just don’t vote for Republicans.

APPENDIX B

Appendix of Chapter 5

B.1 Gold Standard

To further ensure the expertise of my participants, I treated the first of my text snippets as a gold standard and I removed participants that failed to produce a quality translation on the task. Specifically, I modified the MT so that it would always suggest an untranslated version of the term *credit freeze* and then removed participants that failed to reject the suggestion by picking the correct translation ‘gel du crédit‘. The following is the exact gold standard text I included in the study:

> We received your Credit Freeze request and a freeze is now in place on your TransUnion credit report. It will stay in place until you request its removal. A credit freeze prevents lenders from checking your credit report in order to open a new account.

B.2 Statistics

I follow the guidelines of [150] in when reporting of the full linear mixed model outputs in Tables B.1, B.2, and B.3. I only report models that showed statistical significance, leaving out my model for term confidence that showed no significance. I used Satterthwaite’s method to compute p-values for each fixed effect within each table. All p-values are rounded to three decimal places and all other values are rounded to two.

Table B.1: RQ1: Context building time

Likelihood Ratio Test

Effect	logLik	χ^2	*p*
Context Building Time $\sim$ Condition	-346.52	35.09	<**0.001** ***

Random Effects

	Variance	S.D.
Participant (Intercept)	0.41	0.64
Snippet (Intercept)	0.07	0.26
Residual	0.81	0.90

Fixed Effects

	Estimate (β)	SE	95% CI		*t*	*p*
Intercept	0.16	0.17	-0.23	0.56	0.95	0.422
Code of Conduct, 1vs2	-0.31	0.35	-0.98	0.37	-0.89	0.375
Code of Conduct, 1vs3	-0.70	0.35	-1.37	-0.01	-1.99	**0.048** *
Code of Conduct, 1vs4	-0.47	0.37	-1.18	0.25	-1.27	0.205
Marketing, 1vs2	-0.93	0.32	-1.54	-0.32	-2.95	**0.004** **
Marketing, 1vs3	-1.09	0.32	-1.71	-0.46	-3.35	<**0.001** ***
Marketing, 1vs4	-0.57	0.33	-1.20	0.068	-1.72	0.086
Patent, 1vs2	-0.51	0.34	-1.17	0.15	-1.48	0.141
Patent, 1vs3	-1.19	0.33	-1.83	-0.55	-3.57	<**0.001** ***
Patent, 1vs4	-1.00	0.33	CI L	CI U	-3.04	**0.003** **

* p<0.05, ** p<0.01, *** p<0.001

Table B.2: RQ2a: Task time

Likelihood Ratio Test

Effect	logLik	χ^2	*p*
Task Time $\sim$ Context Building Time	-93.26	148.67	<**0.001** ***

Random Effects

	Variance	S.D.
Participant (Intercept)	0.16	0.40
Snippet (Intercept)	0.04	0.20
Residual	0.05	0.23

Fixed Effects

	Estimate (β)	SE	95% CI		*t*	*p*
Intercept	2.37	0.12	2.10	2.65	19.66	<**0.001** ***
Context Building Time	0.27	0.02	0.23	0.31	14.41	<**0.001** ***

* p<0.05, ** p<0.01, *** p<0.001

Table B.3: RQ2b: Quality

Likelihood Ratio Test

Effect	logLik	χ^2	*p*
Quality $\sim$ Term Context Building Time : Snippet	-189.84	13.72	**0.003** **

Random Effects

	Variance	S.D.
Participant (Intercept)	<0.01	<0.01
Snippet (Intercept)	<0.01	<0.01
Residual	0.30	0.54

Fixed Effects

	Estimate (β)	SE	95% CI		*t*	*p*
Intercept	1.09	0.05	0.99	1.18	21.67	<**0.001** ***
TCBT:Code of Conduct	-0.10	0.03	-0.17	-0.04	-3.12	**0.002** **
TCBT:Marketing	-0.03	0.03	-0.09	0.03	-0.88	0.380
TCBT:Patent	0.09	0.03	0.03	0.15	2.77	**0.006** **

* p<0.05, ** p<0.01, *** p<0.001

BIBLIOGRAPHY

[1] Mark S Ackerman, Juri Dachtera, Volkmar Pipek, and Volker Wulf. Sharing knowledge and expertise: The cscw view of knowledge management. *Computer Supported Cooperative Work (CSCW)*, 22(4):531–573, 2013.

[2] Piush Aggarwal, Pranit Chawla, Mithun Das, Punyajoy Saha, Binny Mathew, Torsten Zesch, and Animesh Mukherjee. Hateproof: Are hateful meme detection systems really robust? In *Proceedings of the ACM Web Conference 2023*, pages 3734–3743, 2023.

[3] Syeda Zainab Akbar, Anmol Panda, Divyanshu Kukreti, Azhagu Meena, and Joyojeet Pal. Misinformation as a window into prejudice: Covid-19 and the information environment in india. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–28, 2021.

[4] Rama Akkiraju. Ibm watson tone analyzer–new service now available. *IBM Cloud Blog, Jul*, 16, 2015.

[5] Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Ulrich Germann, Jesús González-Rubio, Robin L Hill, Philipp Koehn, Luis A Leiva, et al. Casmacat: A computer-assisted translation workbench. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 25–28, 2014.

[6] Bethany L Albertson. Dog-whistle politics: Multivocal communication and religious appeals. *Political Behavior*, 37(1):3–26, 2015.

[7] Jeffrey Allen. Post-editing. *Benjamins Translation Library*, 35:297–318, 2003.

[8] Wafa Alorainy, Pete Burnap, Han Liu, and Matthew L Williams. “the enemy among us” detecting cyber hate speech with threats-based othering language embeddings. *ACM Transactions on the Web (TWEB)*, 13(3):1–26, 2019.

[9] Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. Collaborative workflow for crowdsourcing translation. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pages 1191–1194, 2012.

[10] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications, 2016.

[11] Aristotle. *Rhetorica*, page 1:1.

[12] Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In *Companion proceedings of the 2019 world wide web conference*, pages 1100–1105, 2019.

[13] Zahra Ashktorab, Michael Desmond, Josh Andres, Michael Muller, Narendra Nath Joshi, Michelle Brachman, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Christine T Wolf, et al. Ai-assisted human labeling: Batching for efficiency without overreliance. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–27, 2021.

[14] Fatemeh Torabi Asr and Maite Taboada. The data challenge in misinformation detection: Source reputation vs. content veracity. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 10–15, 2018.

[15] Melanie Baak. Racism and othering for south sudanese heritage students in australian schools: Is inclusion possible? *International Journal of Inclusive Education*, 23(2):125–141, 2019.

[16] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[17] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[18] Nicole Baumgarten. Othering practice in a right-wing extremist online forum. *Language@ Internet*, 14(1), 2017.

[19] James Bennett, Stan Lanning, et al. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. New York, 2007.

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

[21] Monika Bickert. Combatting vaccine misinformation. *Facebook*, 2019.

[22] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–19, 2017.

[23] Herbert Bless, Gerald L Clore, Norbert Schwarz, Verena Golisano, Christina Rabe, and Marcus Wölk. Mood and the use of scripts: Does a happy mood really lead to mindlessness? *Journal of personality and social psychology*, 71(4):665, 1996.

[24] Galen V Bodenhausen, Lori A Sheppard, and Geoffrey P Kramer. Negative affect and social judgment: The differential impact of anger and sadness. *European Journal of social psychology*, 24(1):45–62, 1994.

[25] Eduardo Bonilla-Silva. *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States*. Rowman & Littlefield Publishers, 2006.

[26] Lynne Bowker. Fit-for-purpose translation. In *The Routledge handbook of translation and technology*, pages 453–468. Routledge, 2019.

[27] Nadia M Brashier, Emmaline Drew Eliseev, and Elizabeth J Marsh. An initial accuracy focus prevents illusory truth. *Cognition*, 194:104054, 2020.

[28] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1664–1674, 2019.

[29] Alan Brinton. Pathos and the" appeal to emotion": An aristotelian analysis. *History of Philosophy Quarterly*, 5(3):207–219, 1988.

[30] Marc Brysbaert and Michaël Stevens. Power analysis and effect size in mixed effects models: A tutorial. *Journal of cognition*, 1(1), 2018.

[31] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[32] Kenneth Burke. *A grammar of motives*, volume 177. Univ of California Press, 1969.

[33] Pete Burnap and Matthew L Williams. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, 5:1–15, 2016.

[34] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. " hello ai": uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW):1–24, 2019.

[35] Kristie Canegallo. Fighting disinformation across our products. *Google*, 2019.

[36] Robyn Caplan, Lauren Hanson, and Joan Donovan. Dead reckoning navigating content moderation after “fake news”. *Data & Society*, 2018.

[37] Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, 2019.

[38] David L Chen and Raymond J Mooney. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th international conference on Machine learning*, pages 128–135. ACM, 2008.

[39] Quan Ze Chen, Daniel S Weld, and Amy X Zhang. Goldilocks: Consistent crowdsourced scalar annotations with relative uncertainty. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–25, 2021.

[40] Quanze Chen, Jonathan Bragg, Lydia B Chilton, and Dan S Weld. Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[41] Quanze Chen, Jonathan Bragg, Lydia B Chilton, and Daniel S Weld. Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing. *arXiv preprint arXiv:1810.10733*, 2018.

[42] Justin Cheng and Michael S Bernstein. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 600–611, 2015.

[43] Phyllis Pease Chock. The irony of stereotypes: Toward an anthropology of ethnicity. *Cultural anthropology*, 2(3):347–368, 1987.

[44] John J.Y. Chung, Jean Y. Song, Sindhu Kutty, Sungsoo Ray Hong, Juho Kim, and Walter S. Lasecki. Efficient elicitation approaches to estimate collective crowd answers. In *Proceedings of the ACM conference on Computer-Supported Collaborative Work (CSCW '19)*, New York, NY, USA, 2019. ACM.

[45] Adele E Clarke. Situational analyses: Grounded theory mapping after the postmodern turn. *Symbolic interaction*, 26(4):553–576, 2003.

[46] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, pages 1–23, 2019.

[47] John Cook, Stephan Lewandowsky, and Ullrich KH Ecker. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PloS one*, 12(5):e0175799, 2017.

[48] Sven Coppers, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna Van der Lek-Ciudin, Tom Vanallemeersch, and Vincent Vandeghinste. Intellingo: An intelligible translation environment. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[49] Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, 2017.

[50] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[51] Krista De Castella, Craig McGarty, and Luke Musgrove. Fear appeals in political rhetoric about terrorism: An analysis of speeches by australian prime minister howard. *Political Psychology*, 30(1):1–26, 2009.

[52] Nasrin Dehbozorgi, Mary Lou Maher, and Mohsen Dorodchi. Sentiment analysis on conversations in collaborative active learning as an early predictor of performance. In *2020 IEEE Frontiers in Education Conference (FIE)*, pages 1–9. IEEE, 2020.

[53] Richard Delgado and Jean Stefancic. *Critical race theory: An introduction*, volume 87. NyU press, 2023.

[54] Donald DePalma. The language sector in eight charts. *CSA Research*, 2021.

[55] Alain Désilets, Christiane Melançon, Geneviève Patenaude, and Louise Brunette. How translators use tools and resources to resolve translation problems: An ethnographic study. In *Beyond Translation Memories: New Tools for Translators Workshop*, 2009.

[56] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[57] Félix Do Carmo. ‘time is money’and the value of translation. *Translation Spaces*, 9(1):35–57, 2020.

[58] Gavin Doherty, Nikiforos Karamanis, and Saturnino Luz. Collaboration in translation: The impact of increased reach on cross-organisational work. *Computer Supported Cooperative Work (CSCW)*, 21(6):525–554, 2012.

[59] John F Dovidio and Samuel L Gaertner. Aversive racism. 2004.

[60] Ryan Drapeau, Lydia B Chilton, Jonathan Bragg, and Daniel S Weld. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 2016.

[61] Damer T Edward. Attacking faulty reasoning: A practical guide to fallacy-free arguments. *Cengage Learning*, 209, 2008.

[62] Frans H. van Eemeren, Bart Garssen, Erik C. W. Krabbe, A. Francisca Snoeck Henkemans, Bart Verheij, and Jean H. M. Wagemans. Handbook of argumentation theory. 2014.

[63] Elsadig Elsheikh, Basima Sisemore, and Natalie Ramirez Lee. Legalizing othering: The united states of islamophobia. 2017.

[64] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, volume 6, pages 417–422. Citeseer, 2006.

[65] Ethan Fast, Binbin Chen, and Michael S Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657. ACM, 2016.

[66] Marcello Federico, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, Antonio Farina, Domenico Lupinetti, Andrea Martines, et al. The matecat tool. In *COLING (Demos)*, pages 129–132, 2014.

[67] Melissa L Finucane, Ali Alhakami, Paul Slovic, and Stephen M Johnson. The affect heuristic in judgments of risks and benefits. *Journal of behavioral decision making*, 13(1):1–17, 2000.

[68] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

[69] John D Foster. Defending whiteness indirectly: A synthetic approach to race discourse analysis. *Discourse & Society*, 20(6):685–703, 2009.

[70] Roger Fowler. *Language in the News: Discourse and Ideology in the Press*. Routledge, 2013.

[71] Ignacio García. Cloud marketplaces: Procurement of translators in the age of social media. *The Journal of Specialised Translation*, 23:18–38, 2015.

[72] G David Garson et al. Fundamentals of hierarchical linear and multilevel modeling. *Hierarchical linear modeling: Guide and applications*, pages 3–25, 2013.

[73] Roy Gelbard, Roni Ramon-Gonen, Abraham Carmeli, Ran M Bittmann, and Roman Talyansky. Sentiment analysis in organizational work: Towards an ontology of people analytics. *Expert Systems*, 35(5):e12289, 2018.

[74] Sreyan Ghosh, Manan Suri, Purva Chiniya, Utkarsh Tyagi, Sonal Kumar, and Dinesh Manocha. Cosyn: Detecting implicit hate speech in online conversations using a context synergized hyperbolic network. *arXiv preprint arXiv:2303.03387*, 2023.

[75] Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[76] Tarleton Gillespie. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234, 2020.

[77] Ella Glikson and Anita Williams Woolley. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660, 2020.

[78] Amit Goffer and Chaya Zilberstein. Locomotion assisting device and method, January 17 2012. US Patent 8,096,965.

[79] Michael Golebiewski and danah boyd. Data voids: Where missing data can easily be exploited. *New York: Data & Society Research Institute*, 2018.

[80] Robert E Goodin and Michael Saward. Dog whistles and democratic mandates. *The Political Quarterly*, 76(4):471–476, 2005.

[81] Mitchell Gordon, Jeffrey P Bigham, and Walter S Lasecki. Legiontools: a toolkit+ ui for recruiting and routing crowds to synchronous real-time tasks. In *Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pages 81–82. ACM, 2015.

[82] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[83] Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945, 2020.

[84] Colin M Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L Toombs. The dark (patterns) side of ux design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 534. ACM, 2018.

[85] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

[86] Ben Green and Yiling Chen. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–33, 2021.

[87] Spence Green, Jason Chuang, Jeffrey Heer, and Christopher D Manning. Predictive translation memory: A mixed-initiative system for human language translation. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 177–187, 2014.

[88] Spence Green, Jeffrey Heer, and Christopher D Manning. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 439–448, 2013.

[89] Vladas Griskevicius, Noah J Goldstein, Chad R Mortensen, Jill M Sundie, Robert B Cialdini, and Douglas T Kenrick. Fear and loving in las vegas: Evolution, emotion, and persuasion. *Journal of Marketing Research*, 46(3):384–395, 2009.

[90] Ana Guerberof Arenas. What do professional translators think about post-editing. *JoSTrans The journal of specialised translation*, 19:75–95, 2013.

[91] Andrew Guess, Brendan Nyhan, and Jason Reifler. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign. *European Research Council*, 9, 2018.

[92] Danna Gurari and Kristen Grauman. Crowdverge: Predicting if people will agree on the answer to a visual question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3511–3522, 2017.

[93] Jeffry Halverson, Steven Corman, and H Lloyd Goodall. *Master narratives of Islamist extremism*. Springer, 2011.

[94] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*, 2017.

[95] https://lilt.com. The modern way to localize — lilt home. *Lilt*, 2022.

[96] https://memoq.com. memoq — translation and localization management solutions. *MemoQ*, 2022.

[97] https://trados.com. Trados — translation software, cat tool & terminology. *SDL Trados*, 2022.

[98] https://upwork.com. Upwork — the world's work marketplace. *UpWork*, 2022.

[99] Chang Hu, Benjamin B Bederson, and Philip Resnik. Translation by iterative collaboration between monolingual users. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 54–55, 2010.

[100] Chang Hu, Benjamin B Bederson, Philip Resnik, and Yakov Kronrod. Monotrans2: A new human computation system to support monolingual translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1133–1136, 2011.

[101] Fan Huang, Haewoon Kwak, and Jisun An. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*, 2023.

[102] Guoping Huang, Lemao Liu, Xing Wang, Longyue Wang, Huayang Li, Zhaopeng Tu, Chengyan Huang, and Shuming Shi. Transmart: A practical interactive machine translation system. *arXiv preprint arXiv:2105.13072*, 2021.

[103] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. Limitbias! measuring worker biases in the crowdsourced collection of subjective judgments. 2018.

[104] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 407. ACM, 2019.

[105] Jordan S. Huffaker, Jonathan K. Kummerfeld, Walter S. Lasecki, and Mark S. Ackerman. *Crowdsourced Detection of Emotionally Manipulative Language*, page 1–14. Association for Computing Machinery, New York, NY, USA, 2020.

[106] Clayton J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*, 2014.

[107] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67. ACM, 2010.

[108] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635, 2021.

[109] Youxuan Jiang, Catherine Finegan-Dollak, Jonathan K Kummerfeld, and Walter Lasecki. Effective crowdsourcing for a new type of summarization task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 628–633, 2018.

[110] Eric J Johnson and Amos Tversky. Affect, generalization, and the perception of risk. *Journal of personality and social psychology*, 45(1):20, 1983.

[111] Daniel Kahneman. Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, 93(5):1449–1475, 2003.

[112] Sanjay Kairam and Jeffrey Heer. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1637–1648, 2016.

[113] Nikiforos Karamanis, Saturnino Luz, and Gavin Doherty. Translation practice in the workplace: contextual analysis and implications for machine translation. *Machine Translation*, 25(1):35–52, 2011.

[114] Timothy Karr, Craig Aaron, and Free Press. Beyond fixing facebook. *Free Press*, 8, 2019.

[115] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. Improving human-ai partnerships in child welfare: Understanding worker practices, challenges, and desires for algorithmic decision support. In *CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

[116] Brian Felipe Keith Norambuena and Tanushree Mitra. Narrative maps: An algorithmic approach to represent and extract information narratives. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–33, 2021.

[117] Dorothy Kenny, Joss Moorkens, and Félix do Carmo. Fair mt: Towards ethical, sustainable machine translation. *Translation Spaces*, 9:1, 08 2020.

[118] Tuija Kinnunen, Kaisa Koskinen, et al. *Translators' agency*. Tampere University Press, 2010.

[119] Rebecca Knowles, Marina Sanchez-Torron, and Philipp Koehn. A user study of neural interactive translation prediction. *Machine Translation*, 33(1):135–154, 2019.

[120] Zoltán Kövecses. *Where metaphors come from: Reconsidering context in metaphor*. Oxford University Press, USA, 2015.

[121] Anand Kulkarni, Matthew Can, and Björn Hartmann. Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the acm 2012 conference on computer supported cooperative work*, pages 1003–1012. ACM, 2012.

[122] Samuel Läubli, Chantal Amrhein, Patrick Düggelin, Beatriz Gonzalez, Alena Zwahlen, and Martin Volk. Post-editing productivity with neural machine translation: an empirical assessment of speed and quality in the banking and finance domain. *arXiv preprint arXiv:1906.01685*, 2019.

[123] Richard S Lazarus. Cognition and motivation in emotion. *American psychologist*, 46(4):352, 1991.

[124] Matthieu LeBlanc. Translators on translation memory (tm). results of an ethnographic study in three translation services and agencies. *Translation & Interpreting, The*, 5(2):1–13, 2013.

[125] Matthieu LeBlanc. 'i can't get no satisfaction!'should we blame translation technologies or shifting business practices? In *Human issues in translation technology*, pages 63–80. Routledge, 2017.

[126] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[127] Min Kyung Lee and Su Baykal. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*, pages 1035–1048, 2017.

[128] Jennifer S Lerner, Roxana M Gonzalez, Deborah A Small, and Baruch Fischhoff. Effects of fear and anger on perceived risks of terrorism: A national field experiment. *Psychological science*, 14(2):144–150, 2003.

[129] Jennifer S Lerner, Ye Li, Piercarlo Valdesolo, and Karim S Kassam. Emotion and decision making. *Annual review of psychology*, 66, 2015.

[130] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. Assessing the impact of automated suggestions on decision making: Domain experts mediate model errors but take less initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[131] Rebecca Lewis. Alternative influence: Broadcasting the reactionary right on youtube. *New York: Data & Society Research Institute*, 2018.

[132] Rebecca Lewis. Alternative influence: Broadcasting the reactionary right on youtube. 2018.

[133] Daniel J Liebling, Michal Lahav, Abigail Evans, Aaron Donsbach, Jess Holbrook, Boris Smus, and Lindsey Boran. Unmet needs and opportunities for mobile translation ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[134] Cynthia Liem, Markus Langer, Andrew Demetriou, Annemarie MF Hiemstra, Achmadnoer Sukma Wicaksana, Marise Ph Born, and Cornelius J König. Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In *Explainable and interpretable models in computer vision and machine learning*, pages 197–253. Springer, 2018.

[135] Zhuoran Lu and Ming Yin. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[136] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics, 2011.

[137] VK Chaithanya Manam and Alexander J Quinn. Wingit: Efficient refinement of unclear task instructions. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*, 2018.

[138] Beth Innocenti Manolescu. A normative pragmatic perspective on appealing to emotions in argumentation. *Argumentation*, 20(3):327–343, 2006.

[139] Alice Marwick and Rebecca Lewis. Media manipulation and disinformation online. *New York: Data & Society Research Institute*, 2017.

[140] Alice E Marwick. Why do people share fake news? a sociotechnical model of media effects. *Georgetown Law Technology Review*, 2018.

[141] Hana Matatov, Adina Bechhofer, Lora Aroyo, Ofra Amir, and Mor Naaman. Dejavu: A system for journalists to collaboratively address visual misinformation. In *Computation + Journalism Symposium*, Miami, FL, 2018.

[142] Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. Hate begets hate: A temporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–24, 2020.

[143] Arunesh Mathur, Gunes Acar, Michael Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. Dark patterns at scale: Findings from a crawl of 11k shopping websites. *arXiv preprint arXiv:1907.07032*, 2019.

[144] Katerina E Matsa and Shearer Elisa. News use across social media platforms 2018. *Pew Research Center*, Sept 2018.

[145] Priscilla Marie Meddaugh and Jack Kay. Hate speech or "reasonable racism?" the other in stormfront. *Journal of Mass Media Ethics*, 24(4):251–268, 2009.

[146] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th acm international conference on information and knowledge management*, pages 2243–2251, 2018.

[147] David Mellor. Contemporary racism in australia: The experiences of aborigines. *Personality and Social Psychology Bulletin*, 29(4):474–486, 2003.

[148] Tali Mendelberg and John Oleske. Race and public deliberation. *Political Communication*, 17(2):169–191, 2000.

[149] David Merritt, Jasmine Jones, Mark S Ackerman, and Walter S Lasecki. Kurator: Using the crowd to help families with personal curation tasks. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1835–1849. ACM, 2017.

[150] Lotte Meteyard and Robert AI Davies. Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, 112:104092, 2020.

[151] Reine Meylaerts. Translators and (their) norms. *Beyond descriptive translation studies: Investigations in homage to Gideon Toury*, pages 91–102, 2008.

[152] Raphaël Micheli. Emotions as objects of argumentative constructions. *Argumentation*, 24(1):1–17, 2010.

[153] Tanushree Mitra, Clayton J Hutto, and Eric Gilbert. Comparing person-and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1345–1354. ACM, 2015.

[154] Tanushree Mitra, Graham Wright, and Eric Gilbert. Credibility and the dynamics of collective attention. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):80, 2017.

[155] Tanushree Mitra, Graham P Wright, and Eric Gilbert. A parsimonious language model of social media credibility across disparate events. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 126–145. ACM, 2017.

[156] Khouloud Mnassri, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. Hate speech and offensive language detection using an emotion-aware shared encoder. *arXiv preprint arXiv:2302.08777*, 2023.

[157] Saif M Mohammad and Peter D Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics, 2010.

[158] Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[159] Joss Moorkens and Sharon O'Brien. Post-editing evaluations: Trade-offs between novice and professional participants. In *Proceedings of the 18th annual conference of the European association for machine translation*, 2015.

[160] Joss Moorkens and Sharon O'Brien. Assessing user interface needs of post-editors of machine translation. *Human issues in translation technology*, pages 109–130, 2017.

[161] Christopher H Morrell. Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics*, pages 1560–1568, 1998.

[162] Adam Mosseri. Addressing hoaxes and fake news. *Facebook*, 2016.

[163] Bonnie M Muir. Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies*, 27(5-6):527–539, 1987.

[164] Tanja Munz, Dirk Väth, Paul Kuznecov, Thang Vu, and Daniel Weiskopf. Visual-interactive neural machine translation. In *Graphics Interface 2021*, 2021.

[165] Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*, 2019.

[166] Brian Felipe Keith Norambuena, Tanushree Mitra, and Chris North. Narrative sensemaking: strategies for narrative maps construction. In *2021 IEEE Visualization Conference (VIS)*, pages 181–185. IEEE, 2021.

[167] Christiane Nord. Functionalist approaches. *Handbook of translation studies*, 1:120–128, 2010.

[168] Ihudiya Finda Ogbonnaya-Ogburu, Angela DR Smith, Alexandra To, and Kentaro Toyama. Critical race theory for hci. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–16, 2020.

[169] Charles Kay Ogden and Ivor Armstrong Richards. *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*, volume 29. K. Paul, Trench, Trubner & Company, Limited, 1927.

[170] John F Padgett and Christopher K Ansell. Robust action and the rise of the medici, 1400-1434. *American journal of sociology*, 98(6):1259–1319, 1993.

[171] Gordon Pennycook and David G Rand. The implied truth effect: Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings. 2017.

[172] Gordon Pennycook and David G Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 2018.

[173] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*, 2017.

[174] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52, 2021.

[175] Alexander J Quinn and Benjamin B Bederson. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1403–1412. ACM, 2011.

[176] Meena Rambocas and Barney G Pacheco. Online sentiment analysis in marketing research: a review. *Journal of Research in Interactive Marketing*, 12(2):146–163, 2018.

[177] Flávio Ribeiro, Dinei Florencio, and Vítor Nascimento. Crowdsourcing subjective image quality evaluation. In *2011 18th IEEE International Conference on Image Processing*, pages 3097–3100. IEEE, 2011.

[178] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 131–141, 2020.

[179] Hanna Risku, Regina Rogl, and Jelena Milosevic. Translation practice in the field: Current research on socio-cognitive processes. *Translation Spaces*, 6(1):3–26, 2017.

[180] Hanna Risku and Florian Windhager. Extended translation: A sociocognitive research agenda. *Target. International Journal of Translation Studies*, 25(1):33–45, 2013.

[181] Sarah T Roberts. *Behind the screen*. Yale University Press, 2019.

[182] Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518, 2017.

[183] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[184] Sebastin Santy, Kalika Bali, Monojit Choudhury, Sandipan Dandapat, Tanuja Ganu, Anurag Shukla, Jahanvi Shah, and Vivek Seshadri. Language translation as a socio-technical system: Case-studies of mixed-initiative interactions. In *ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 156–172, 2021.

[185] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, 2019.

[186] Mike Schaekermann, Carrie J Cai, Abigail E Huang, and Rory Sayres. Expert discussions improve comprehension of difficult cases in medical image assessment. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13, 2020.

[187] Jordan Scott. Does racism equal prejudice plus power? *Analysis*, 82(3):455–463, 2022.

[188] Elizabeth A Segal. *Social Empathy: The Art of Understanding Others*. Columbia University Press, 2018.

[189] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[190] Craig A Smith and Phoebe C Ellsworth. Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4):813, 1985.

[191] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics, 2008.

[192] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[193] Melodie Yun-Ju Song and Anatoliy Gruzd. Examining sentiments and popularity of pro-and anti-vaccination videos on youtube. In *Proceedings of the 8th International Conference on Social Media & Society*, page 17. ACM, 2017.

[194] Ivan Srba, Gabriele Lenzini, Matus Pikuliak, and Samuel Pecar. Addressing hate speech with data science: an overview from computer science perspective. *Hate Speech-Multidisziplinäre Analysen und Handlungsoptionen*, pages 317–336, 2021.

[195] Kate Starbird. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[196] Kate Starbird, Ahmer Arif, and Tom Wilson. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, 2019.

[197] Luc Steels and Manfred Hild. *Language grounding in robots*. Springer Science & Business Media, 2012.

[198] Walter G Stephan, Oscar Ybarra, and Kimberly Rios. Intergroup threat theory. 2016.

[199] Carlo Strapparava, Alessandro Valitutti, et al. Wordnet affect: an affective extension of wordnet. In *Lrec*, volume 4, page 40. Citeseer, 2004.

[200] Anselm Strauss and Juliet Corbin. Grounded theory methodology: An overview. 1994.

[201] Derald Wing Sue. *Microaggressions in everyday life: Race, gender, and sexual orientation*. John Wiley & Sons, 2010.

[202] Derald Wing Sue, Christina M Capodilupo, Gina C Torino, Jennifer M Bucceri, Aisha Holder, Kevin L Nadal, and Marta Esquilin. Racial microaggressions in everyday life: implications for clinical practice. *American psychologist*, 62(4):271, 2007.

[203] Carlos SC Teixeira and Sharon O'Brien. Investigating the cognitive ergonomic aspects of translation tools in a workplace setting. *Translation Spaces*, 6(1):79–103, 2017.

[204] Larissa Z Tiedens and Susan Linton. Judgment under emotional certainty and uncertainty: the effects of specific emotions on information processing. *Journal of personality and social psychology*, 81(6):973, 2001.

[205] Charles Tilly. Survey article: power—top down and bottom up. *Journal of Political Philosophy*, 7(3):330–352, 1999.

[206] Alexandra To, Wenxia Sweeney, Jessica Hammer, and Geoff Kaufman. " they just don't get it": Towards social technologies for coping with interpersonal racism. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–29, 2020.

[207] Iulianna Van der Lek-Ciudin, Tom Vanallemeersch, and Ken De Wachter. Contextual inquiries at translators' workplaces. *Proceedings of the TAO-CAT, Angers, France*, pages 18–20, 2015.

[208] Teun A Van Dijk. *Racism and the Press*. Routledge, 2015.

[209] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[210] Andy Way. Traditional and emerging use-cases for machine translation. *Proceedings of Translating and the Computer*, 35:12, 2013.

[211] David Gray Widder, Laura Dabbish, James D Herbsleb, Alexandra Holloway, and Scott Davidoff. Trust in collaborative automation in high stakes software engineering work: A case study at nasa. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[212] Timothy D Wilson and Nancy Brekke. Mental contamination and mental correction: unwanted influences on judgments and evaluations. *Psychological bulletin*, 116(1):117, 1994.

[213] Dennis Wixon, Karen Holtzblatt, and Stephen Knox. Contextual design: An emergent view of system design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '90, page 329–336, New York, NY, USA, 1990. Association for Computing Machinery.

[214] Stefan Wojcik, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, MB Hunzaker, Keith Coleman, and Jay Baxter. Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. *arXiv preprint arXiv:2210.15723*, 2022.

[215] Meng-Han Wu and Alexander James Quinn. Confusing the crowd: Task instruction quality on amazon mechanical turk. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*, 2017.

[216] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399, 2017.

[217] Yan Xia, Haiyi Zhu, Tun Lu, Peng Zhang, and Ning Gu. Exploring antecedents and consequences of toxicity in online discussions: A case study on reddit. *Proceedings of the ACM on Human-computer Interaction*, 4(CSCW2):1–23, 2020.

[218] Anbang Xu and Brian Bailey. A reference-based scoring model for increasing the findability of promising ideas in innovation pipelines. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1183–1186. ACM, 2012.

[219] Qian Yang, Aaron Steinfeld, and John Zimmerman. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

[220] Jason R Young. The role of fear in agenda setting by television news. *American Behavioral Scientist*, 46(12):1673–1695, 2003.

[221] Muhammad Bilal Zafar, Parantapa Bhattacharya, Niloy Ganguly, Saptarshi Ghosh, and Krishna P Gummadi. On the wisdom of experts vs. crowds: discovering trustworthy topical news in microblogs. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 438–451. ACM, 2016.

[222] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*, 2018.

[223] Jost Zetzsche. Freelance translators' perspectives. In *The Routledge handbook of translation and technology*, pages 166–182. Routledge, 2019.

[224] Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, et al. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 603–612. International World Wide Web Conferences Steering Committee, 2018.

[225] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.

[226] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.