Accounting for Offensive Speech as a Practice of Resistance

Mark Díaz*
Google Research
markdiaz@google.com

Razvan Amironesei*
Google Research
amironesei@gmail.com

Laura Weidinger
DeepMind
lweidinger@deepmind.com

Iason Gabriel
DeepMind
iason@deepmind.com

Abstract

Tasks such as toxicity detection, hate speech detection, and online harassment detection have been developed for identifying interactions involving offensive speech. In this work we articulate the need for a relational understanding of offensiveness to help distinguish denotative offensive speech from offensive speech serving as a mechanism through which marginalized communities resist oppressive social norms. Using examples from the queer community, we argue that evaluations of offensive speech must focus on the impacts of language use. We motivate this use of language in Cynic philosophy and use it to frame a use of offensive speech as a practice of resistance. We also explore the degree to which NLP systems may encounter limits to modeling relational context.

1 Introduction

Tasks such as the detection of toxicity, hate speech, and online harassment have been developed to identify and intervene in situations that have the potential to cause significant social harm. These tasks for identifying and classifying offensive or undesirable language have gone by different names (see: (Waseem et al., 2017; Balayn et al., 2021)) and have employed varying task definitions, but they are united by a goal of reducing harm and breakdowns in civil discourse. Because language use varies contextually, it is difficult to model the nuanced social context that informs whether language produces harm. Offensive language classification and related tasks capture different forms of undesirable language, such as language that is rude, incites hate, causes offense, or causes people to disengage from online interaction.

In this paper, we discuss a form of offensive language that has not previously received much research attention in the machine learning (ML) community, namely offensive language that is beneficial in its use and whose prosocial effects are sociologically and historically documented. In other words, language that uses terminology which is often noted as offensive, but which is not perceived as offensive in particular contexts of use. We distinguish this language with contextually-specific beneficial impacts as *reappropriated*. Understanding how to characterize and model this kind of language is important not only because of its widespread use, but also because of the critical sociological role it can play, particularly within marginalized communities.

We contribute: 1) a framing of offensiveness that accounts for socially productive uses of denotatively offensive language; 2) a general understanding of offensive language that builds from definitions of hate speech and toxicity to show the difficulty of operationalizing relational context; 3) specific challenges and directions for improving how we operationalize relational aspects of offensiveness.

We also call on researchers developing offensive speech classification tasks to engage with offensiveness as a social relation that arises not just between individuals in communication but also between communities of discourse. In other words, offensiveness and its impacts are best understood from the perspectives of those already embedded in relationships that structure who produces, receives, perceives and who is targeted (i.e., who is implicitly or explicitly named) by offensive speech. We separately name reception and perception to distinguish between the recipient(s) of a message, such as an alias tagged in a Tweet, and those who may not be the intended audience but to whom the content is visible. However, practical challenges in data collection and annotation task design can cause offensiveness to be implicitly operationalized as a semantic property of language.

Our discussion begins with the foundations of

*Authors contributed equally

Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 192 - 202

offensiveness, adding to challenges that have been highlighted by others. We also provide considerations and steps forward for improving automated offensiveness classification. We argue that accounting for the social and historical constitution of offensive language is important for the responsible development of automated and semi-automated tools to identify offensive language. To do so, in turn, deepens our understanding of offensive speech and recognizes the disparate impacts or ends of offensive language (e.g., as a means to silence others; as a means to challenge existing social structures).

2 A Working Definition of Offensiveness

Natural language processing (NLP) systems have historically faced challenges identifying and classifying denotatively offensive language used with inoffensive connotation (Ashwitha et al., 2021; Weitzel et al., 2016; Sun et al., 2021). A challenge inherent to defining different forms of hateful, toxic, or offensive language is that the characterization of these terms is necessarily socially, culturally, and politically specific. For this reason, Hovy and Yang (2021) identify the robust inclusion of social context in understanding language as a key missing component for the success of modeling approaches. Building from definitions of hate speech and toxicity, we establish a working definition of offensiveness that can better account for missing social context.

Hate speech is typically defined to link the use of derogatory language to a person or people based on group membership, such as "some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics" (Basile et al., 2019). One of the key characteristics of this definition is that it focuses on the injury applied to a specific subject of offense. Broadly, definitions of hate speech underscore a need to identify who the target is in order to assess offense or harm. We argue that the target of offensive language is best understood with respect to differential relations and power dynamics between them and the producers, receivers, and perceivers of offensive speech.

In contrast to hate speech and other definitions of offensive language, toxicity is specifically oriented around the measurable outcome of language use. Defined as, "a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion,"[1] it does not engage explicitly with injury or harm, but links it to measurable behavior. Toxicity helps bring attention to the ends of reappropriated offensive language, and what characteristics can help distinguish it from denotative uses of offensive language. On its own, however, it is not clear that it suffices to protect the user's best interest, as it does not engage explicitly with the subject of verbal injury. The definition serves those with an interest to maximize user engagement, i.e. to minimize the chances of the user "leaving a discussion". It is conceivable that users who experience offense may not leave a discussion, for example in cases where users are habitually exposed to the relevant offense, such as microaggressions.

Hate speech and toxicity help us to construct a relational view for grounding a more robust definition of offensiveness– that is, a view that considers the social relations among targets of offensive language, and the producers, receivers, and perceivers (each of whom can also be a target) of denotatively offensive speech. This view is focused on identifying the network of social relations rather than some essential attribute of words or phrases.

3 Drag Queens and the Cynic Perspective

In this section, we analyze language use within the queer community to show how social relations and the ends of targeted language can help us to understand how denotatively offensive language 1) can have expressly beneficial ends and 2) can work as a social practice of collective resistance which fulfills a function of "self-innoculation" against out-group antagonism. We specifically analyze mock impoliteness used by drag queens that implicitly offers a critique of exclusionary sexual mores and social attitudes that hinder the self-expression of queer communities. These social attitudes include the demonization of queer sexuality, gender expression and visibility. We foreground our example with the case of Cynic offensive speech, which we present as a historical practice of resistance to unreflective social conventions. We show that denotatively offensive language as vehiculated by drag queens is not an isolated sociological phenomenon but instead should be inscribed in a history of practices of resistance.

[1] https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages: :text=%EF%BB%BFAttributestext=Perspective's%20main%20attribute%20is%20TOXICITY,make%20you%20leave%20a%20discussion%E2%80%9D

Culpeper (2011) defines mock impoliteness as language which "consists of impolite forms whose effects are (at least theoretically for the most part) canceled by the context." The author continues "[...] mock impoliteness in theoretical terms [is understood] as involving the canceling of impoliteness' perlocutionary effects flowing from a conventionalised impoliteness formula when an obvious mismatch emerges with the context it is used in." One key aspect that we want to emphasize here is how mock impoliteness works to both reduce the harmful effects of targeted insult while supporting social bonding and relationship building.

3.1 Mock Impoliteness in the Queer Community

The use of mock impoliteness is not exclusive to the queer community. However, compared with other uses of mock impoliteness, its use in the queer community acts as a form of social resistance (McKinnon, 2017). Although, building in-group solidarity and social bonding are significant effects of mock impoliteness, we focus on what it means for queer individuals to "self-innoculate" against offensive language by practicing employing it themselves. McKinnon (2017) uses mock impoliteness to show that "utterances, which could potentially be evaluated as genuine impoliteness outside of the appropriate context, are positively evaluated by in-group members who recognize the importance of "building a thick skin" to face a hostile environment both explicitly via the deployment of offensive language which targets marginalized communities (e.g., slurs) and implicitly via the structures of civil language, which may inhibit certain forms of expression by marginalized communities (e.g., comments highlighting nontraditional gender expression). One complexity in interpreting language that relies heavily on context, is the "context collapse" that takes place on platforms such as Twitter (Marwick and Boyd, 2011). A drag queen may target the drag community in a Tweet, relying on shared contextual markers with their targeted audience - for example, assuming shared norms of mock impoliteness, and a mutual understanding that slurs are not intended literally. However on platforms such as Twitter, where the wider public can see these messages, perceivers may lack this context and thus interpret such utterances as offensive. This "context collapse" must be accounted for to assess the content of a given Tweet.

(a)

(b)

Figure 1: Examples of mock impoliteness from prominent drag queens featured on RuPaul's Drag Race.

Oliva et al. (2021) describe erroneously high toxicity probabilities (provided by the Perspective API) for language from drag queens on Twitter[2]. Although toxicity is distinct from offensiveness, their analysis shows how the difficulty of implementing a relational approach and detecting relational context in practice can cause the concept to be misapplied. This difficulty applies not only to classifying toxicity, but also classifying hate speech, offensiveness, and other language rooted in relational context. The authors compare tweets that contain queer vernacular produced by drag queens to racist tweets written by white supremacists that are predicted to have low toxicity. As the authors point out, many of the swear words and slurs used among drag queens that might otherwise be considered insulting or rude, are used playfully. To characterize the constructive nature of language in this example, it is not only important to understand the individual relation and the type of humor between two drag queens engaging with each other, but also to understand their marginalized social positions relative to broader society. This example brings to light the difficulty of not only developing a relational framing but also practical challenges in identifying this relational context in data.

[2]Oliva et al. (2021) offer an example tweet of mock impoliteness: tweet by drag queen Darienne Lake, "So proud of this bitch. Love seeing you on @AmericanIdol."

Contrary to Oliva et al., we are not focused here on toxicity. Rather, we assess the above example in terms of offensiveness and disambiguate between hate speech, toxicity and offensive speech not only at the definitional level but also at a conceptual level. For example, applying Perspective API's definition of toxicity is complicated by the fact that, presented without context to a perceiver or data annotator, this language may be indistinguishable from rude, uncivil, indecent and impolite exchange. In addition, with context but without knowledge of sociological norms within the community, it may still be assumed by a perceiver that the exchange also leads to disengagement, thus qualifying the interaction as toxic. This challenge for outside perceivers highlights a need to identify which context is required in instances where members of a community of discourse - in this case drag queens - may break normative rules of civility (such as by using widely acknowledged hateful terms like "fag", "tranny", or "dyke") and consensually use language deemed uncivil by mainstream standards.

3.2 Mutual Recognition and Consent

It is important to emphasize that social positionality, (that is, the roles one can fulfill in a given social context) is crucial for disambiguating offensive content. Take the example of the Cynics to whom offense fulfills an ascetic role as part of a larger project of living one's life by practicing spiritual exercises. Cynic philosophy is defined by a denunciation of normative social conventions and by a demand to "return to a simple life in conformity with nature" (Hadot, 2002). Cynic philosophy, as set out by the philosopher Diogenes, established ethical practices that its proponents put forward to support a virtuous way of life, which was achieved through severe self-discipline and the strategic use of offensive language. Similar to mock impoliteness, cynic insult was used as a way to critique unreflective social norms from a position of subjugation. In this respect they both play a sociological role as practices of resistance. However, whereas the outcome of offensive language for Cynics was to further the way of life of an ancient school of thought, for drag queens, the outcome is solidaristic bonding, identity formation, and queer survival in the face of marginalization.

Thus, we must ask if consenting to being both the producer and recipient of offensive language is a sufficient condition for an observer to identify the language as inoffensive or for a platform to tolerate it. In this case, the drag queens directing tweets at each other recognize a shared queer identity referenced in a specific type of offensive language and can be reasonably confident that their messages will be considered a practice of mock impoliteness based on queer vernacular and social dynamics. Critically, mock impoliteness as used by drag queens and members of the queer community features slurs and insults that have precisely been used in targeted harassment against them. Among members of the queer community, interactions such as these are predicated upon a dynamic of group exchanges, recognition of group membership, and awareness of such language as used by outgroup members (e.g., the use of slurs such as 'faggot' or 'tranny' and insults focused on sexual behavior and femininity).

Recognition of the sociological norms of mock impoliteness is reflected in reciprocal, consensual engagement. In this way, mock impoliteness is a mutually socially constructed phenomenon. However, it is critical to note that language use within groups may not be consistently used or recognized by all. For example, someone new to the drag community might be initially unfamiliar with mock impoliteness, mistaking it for malice, and others within the community may simply not engage in mock impoliteness. In these cases, reciprocal, consensual engagement is not achieved. This shows that recognition of offensive discourse is still an insufficient condition to identify mock impoliteness and that it may need to be followed by explicit consent to offensive language.

Although we name the combination of recognition and mutual consent to offensive language as reasons for potentially tolerating its use on a platform, we do not argue in favor of protecting hateful exchanges among producers and receivers who target individuals or groups outside of their own. This would include, for example, two homophobic individuals bonding over and consenting to uses of queer slurs among themselves. Any community of discourse that promotes racist, homophobic, xenophobic, or similar ideas should be submitted to scrutiny. Out of context, drag queen's reappropriated offensive speech may appear to do just this, showing that recognition and consent between producer and recipient to offensive language are not sufficient conditions for identifying mock impoliteness. To build from linguistic recognition

and consent, we ask why and, in particular, to what ends do drag queens invoke offensive language on the platform?

As Oliva et al. describe, much of the language used in the tweets they analyzed is reflective of mock impoliteness among members of the queer community, which is a critically important act of socializing– or training– ingroup members to inure and defend themselves from derogatory remarks lobbed by outgroup members. While mock impoliteness features swear words and often employs homophobic slurs, this language can be distinguished from offensive language invoked by outgroup harassers in at least two important ways.

First, mock impoliteness serves to inure queer individuals to homophobic attacks from outgroup members, as well as to hostility from other community members (McKinnon, 2017; Murray, 1979). In this way, offensiveness is used as a rhetorical means to prosocial ends that effectively resist to exclusionary social norms that seek to define queer existence. Taking automated action on offensiveness as an end in itself to be identified stands to ignore its prosocial impacts and ignores the web of social relations that support queer survival. In order to develop content moderation and related processes that predict and mitigate harm, offensiveness must also be conceptualized in such a way that accounts for the beneficial impacts, or ends, of language rather than treating offensiveness as a fixed, negative property of language itself. While mock impoliteness might be read as targeted harassment by a random or non-queer audience, a relational lens makes clear that offensiveness can operate within a dynamic relation of in-group recognition that consolidates the formation of a social identity.

In this respect, the queer practice of mock impoliteness parallels a Cynic practice of training for adversity. The Cynic analogy of the adaptive mouse describes the actions of a mouse adapting itself to harsh living environments. The analogy describes how Diogenes, the preeminent Cynic, rolled himself "over hot sand, while in winter he used to embrace statues covered with snow, using every means of inuring himself to hardship." The practice of inuring oneself to hardship is consistent with the description of the drag queens practice of "building a thick skin." The Cynics and drag queens have distinct motivations and engage in distinct behavior, however queer practices of self-inoculation parallel a lineage of "building a thick skin", figuratively and literally within a history of practices of training for adversity.

Second, as the case of mock impoliteness illustrates, characterizing the ends (or potential ends) of offensiveness is central. The goal of this work is not to propose a universal taxonomy of offensiveness. Instead, we turn attention to understanding how offensiveness functions as a social practice and the ends it produces in order for us to account for the various ways of operationalizing it. Focusing on the complex intentions of offensiveness allows us to describe its relational nature.

We characterize two types of ends – those that are individual and those that are plural or collective and related to belonging to a recognized social identity. Individual ends refer to the impacts local to a specific interaction and the people directly engaged. This includes the interlocutors as well as individual entities named explicitly or implicitly in an utterance. Plural ends refer to those that impact individuals who are not specifically named or involved but who may bear witness to an interaction describing their social group, such as through a curated social media feed. This includes utterances that name groups or communities. Though they don't specifically discuss the ends of offensiveness or abusive language (Waseem et al., 2017) importantly highlight that many classification tasks can be understood in relation to whether they focus on language that is directed toward a specific individual or entity or whether they focus on language directed toward a generalized group. In discussing ends we posit that intergroup linguistic based recognition between parties are necessary but insufficient conditions for identifying offensive language with beneficial outcomes. A sufficient condition for allowing the presence of offensive language is the verifiable absence of harm.

But our argument with respect to the articulation of ends goes further. As a way to illustrate the sociological relevance of mock impoliteness in the context of drag queen discourse as a practice of "building a thick skin," we placed the concept in a history of ethical and spiritual practice of offensive language in the Cynic philosophy.

In both the case of mock impoliteness and the case of the Cynics there are distinct yet connected practices of engaging with adversarial conditions and social norms. The Cynic use of offensive speech is focused on creating a space for the practice of a virtuous way of life actualized via sys-

tematic practices of training and endurance. Offense as a public act of provocation instantiates the ethical substance of a "life of battle and struggle against and for others (Foucault and Foucault, 2012)." queer practices of building thick skin via offensive speech is similarly defined by the preservation and self-expression of a social identity. In this way, queer mock impoliteness also stands as an example of reclaiming offensive language, which has been studied and documented in relation to social justice targets, such as misogyny (Gaucher et al., 2015) and ableism (Smith, 2012).

4 Rethinking Offensiveness for Machine Learning Practice

In our assessment of mock impoliteness, we motivate a relational frame with an example of marginal discourse to bring attention to social relations reflective of power dynamics in society. However this does not emerge from a void. Rather, it is grounded in a broad range of conceptual precedents that articulate the social, ethical and epistemological role of relationality. For example, Foucault's analyzes power (pouvoir) as a techno-social relation of subjection (Foucault, 2012) and of freedom (Foucault, 1982), Arendt theorizes power as a capacity to act in concert with others (Arendt, 2013), Weber understands power (Macht) as the exercise of a will on another will (Weber, 2019), and Patricia Hill Collins' analysis of lived experience within the domain of Black feminist epistemologies are key to our argument (Collins, 2002). In particular, Hill Collins analyzes the "connections between knowledge and power relations" and the particular forms of knowledge operative in Black women's lived experience.

In all of these cases, relationality unveils and constitutes new forms of knowledge, new forms of ethical action, and new forms of individual and collective experience. More specifically, in the space of AI ethics, Birhane (2021) discusses relational ethics as a "framework" that re-examines "hierarchical power asymmetries," how the "contingent and interconnected background that algorithmic systems emerge from (and are deployed to) in the process of protecting the welfare of the most vulnerable." Cooper et al. (2022) put forward a framework for relational accountability, Viljoen (2021) proposes a relational theory of data governance which shows how "data relations result in supra-individual legal interests" that in turn "materialize unjust social relations" via data flows which order in particular ways social existence.

At the conceptual level a relational frame shows that particular social and historical context is crucial to account for how offensive speech emerges and is constituted in a network of social relations by introducing discursive markers of style (as our example of mock impoliteness in drag queen shows) rather than relying on the perceived essential attribute of words or phrases. It also lays the foundation for practical experimentation and highlights avenues for designing classification tasks and identifying what context must be accounted for in data annotation, dataset construction, and modeling techniques.

4.1 Providing Context to Annotators

Drawing from a relational frame, there are opportunities to improve annotation task design and data collection by leveraging intuitive human understandings of social context. For example, annotation task instructions can invoke relational context that humans implicitly use to judge the offensive nature of language. Sap et al. (2019) introduced dialect priming to annotators as a contextual cue to the origins of an utterance. They primed annotators with a measure of an utterances' alignment with AAE, which significantly reduced the degree to which they rated AAE utterances as toxic. In addition, when asked to consider the tweet author's likely race, annotators were also less likely to identify AAE tweets as toxic.

Identifying an utterance as in alignment with AAE implicitly introduces sociologically informed norms about language use, the contexts in which it is likely to be used and consented to, and broader social context about how language produced by the likely author may be perceived. However, other work has found limited success in providing annotators with more context (Pavlopoulos et al., 2020). More work is needed to explore how exactly annotators use additional context, and in which instances additional context is most influential it is not clear exactly how annotators used contextual information to make sense of the tweet prompts.

Moreover, research is needed to explore the role of context in data annotation, for example, to explore avenues of capturing why a rater may annotate utterances the way that they do. Annotators' sense making practices– how they rely on different contextual clues when making judgments– remain

generally unclear in text labeling. Significant correlations have been shown between annotators' political views and their ratings of antiblack speech, suggesting that political viewpoints may also be worth considering or documenting when selecting annotators or evaluating interrater agreement (Sap et al., 2021). Similarly, Prabhakaran et al. (2021) found differences between black and non-black annotator's labels of sentiment on age-related text prompts. Conversely, little work has explored how raters fill in contextual gaps when details are not provided, for example what kinds of assumptions might be made about the producer of an utterance, which, in terms of race, critical race scholars would suggest Whiteness is assumed (Sue, 2006). Given the use of the globally distributed crowd workforce, the ways in which these assumptions may differ across regions also stands to be explored.

A parallel direction to highlighting additional social context is to select data annotators with the ability to recognize the sociological norms embedded in context. The recognition of the norms surrounding language use is predicated on knowledge of, experience with, or proximity to specific forms of language use. In this vein, machine learning researchers have highlighted a need for considering annotator social identity in both dataset documentation (Prabhakaran et al., 2021; Díaz et al., 2022) as well as in modeling techniques (Davani et al., 2021). In Sap et al. (2019)'s work it is not clear whether and how individual annotators may have taken race and dialect information into account when making judgments.

However, prior work on data annotation by Patton et al. (2019) shows that annotators' social identity and lived experiences can shape the cues they draw from when making annotation judgments. Moreover, they demonstrate that lived experience can inform different judgments in comparison with annotators who have been formally educated and trained on the concepts being annotated. This has particularly important implications for the annotation of linguistic phenomena such as mock impoliteness by drag queens and others in the queer community, which may not be familiar or legible to annotators who do not share a queer social identity.

Mock impoliteness and the Cynic thought are key to unveiling historical and sociological reasons for offensiveness language use. We rely on the intuition that annotators with knowledge and/or shared group membership are more likely to be exposed to sociological norms used within their own social groups than to norms in other groups. For members of marginalized groups, recognizing these sociological norms is also an implicit recognition of the social and ethical modes of resistance that they represent and embody. At the same time, it is critical to acknowledge that no social group is monolithic, so there are inherent limits to both the degree to which members are representative of other members as well as the degree to which they can be expected to recognize mock impoliteness from other in-group members.

4.2 Ends and Outcomes of Offensiveness

Another avenue for improving offensiveness classification is to bring its measurable outcomes into focus. Treating offensive language as a means to an end underscores decisions about which of its different ends we seek to identify, whether beneficial in the case of queer resistance or more negative in cases of insult. Identifying which impacts to focus on can shift design and implementation practices.

By bringing focus to the likelihood of offensive language to cause a person to disengage from interaction, toxicity as defined by Jigsaw provides an example of incorporating measurable ends into the definition of the classification task. As such, the definition inspires specific behaviors or outcomes to be measured. Acknowledging the ways in which members of marginalized communities stand to be disproportionately harmed, it is also prudent to consider how people with marginalized identities may respond to negative or harmful language differently from others. For example, because they are more likely to endure disrespectful or harassing language, people with marginalized identities may endure offensive language in interactions longer than others. This means that a behavioral measure, such as exiting a conversation, may be differently predictive of offensiveness depending on the social identities of targets involved.

Nonetheless, capturing offensive language under the label 'toxicity' is an interesting departure from other labels used to describe offensiveness, such as 'misogynous' or 'aggressive', because of its focus on using observable behavior as a metric. Mishra et al. (2019) also take into account observable behavior by modeling sexist and racist tweets using author profiles. Doing so captured repeated behaviors and hateful discourses represented in certain profiles and improved model performance. Model-

ing user profile discourse stands as a way to combine language modeling with measurable behavior for identifying offensive language. As Cheng et al. (2017) show, trolling behavior can be predicted, in part, from user mood as measured through recent user history, which offers a signal of unwanted speech. . Of course, not all offensive language is produced by online trolls and not all online trolls produce offensive language. However, as Mishra et al. show, capturing histories of behavior and interactions, including their associated norms and patterns of speech can be an avenue for using behavioral measures to improve language modeling.

Bringing focus to platform and account-level behavior also affords the ability to infer additional social context, such as user political alignment, which scholars have predicted based on interactions and follower lists on Twitter (Colleoni et al., 2014). In addition, Hovy (2015) found that training gender- and age-"aware" classifiers using embeddings created from user reviews filtered by author age and gender provided modest, consistent improvements in topic classification and sentiment analysis tasks. Using similar techniques, information about content producers might be used to prevent their content from over-moderation, which Oliva et al. suggest impacts drag queens on Twitter. As yet another alternative, modeling discourse and discursive styles might be used to allow perceivers to selectively filter out undesired language from their social media feeds. At the platform level, this raises the potential for serious privacy and surveillance concerns which must be considered. However, even at the level of individual interactions, work in NLP has shown it is possible to identify aspects of interpersonal communication in chat contexts, such as whether a relation is cooperative (Rashid et al., 2020).

Operationalizing offensiveness in terms of specific ends also allows developers to focus on particular harms and, for content moderation, add specificity to whether individual or symbolic harms may be at stake. On platforms where individual interactions may be visible to many, such as Twitter, there are murky questions about how to prioritize impacts on targets in comparison to impacts on perceivers if and when they diverge. For example, a digital passerby (perceiver) who is unfamiliar with queer mock impoliteness may take offense based on language used within a tweet, even if the tweet producer and receiver are mutually engaging in mock impoliteness. More broadly, misunderstandings of mock impoliteness might normalize offensive language use for those who do not recognize its contextual nature. Platforms must consider if and in which circumstances impacts to perceivers, who may interpret a message as earnest, warrant consideration over the target of the message.

These considerations are particularly salient in regard to platforms that allow one-to-many communication, however they are also relevant for platforms or spaces limited to private or one-on-one interactions. Certain kinds of language that we have not discussed here, for example antisemitic utterances, may be harmful even if the recipient is not offended and no third party reads the content (e.g., if the content is shared in a private message). As such messages can incite violence or hate by asserting that some groups or individuals are of lower value than others, these messages can cause harm. This may warrant their detection and prohibition on a given platform.

4.3 Limits to Modeling Context

Identifying context and the ends of offensiveness as key components for defining offensiveness raises challenges and illustrates limits to developing automated classification tasks. One major difficulty lies in inferring or recording contextual data.

Social identity has been a through line in our examples of the role context plays in both how offensiveness operates as well as in machine learning annotation. However, annotating or documenting social identity, in particular, becomes challenging and ethically dubious, as Scheuerman et al. (2020) discuss with regard to gender and race annotation for computer vision. NLP techniques that have been used to infer or extract demographic characteristics such as age (Hovy and Søgaard, 2015), can provide helpful approximations; however they are limited by similar ethical concerns to annotation. Moreover, identifying social characteristics of an individual or group targeted in, receiving, or perceiving an utterance may be impossible to determine. Documenting demographic information of data annotators and explicitly inviting reflection on identity to broaden the sociological norms a rater pool is able to recognize may be a promising alternative. Importantly, this requires limits to protect the privacy of workers, particularly if workers with marginalized identities are repeatedly sought out for their ability to recognize how people in their

communities communicate. This is to say that evaluating the social identities of actors involved in an online interaction, as well as the social identities of those perceiving or annotating the interaction, is a hurdle. Indeed, social media platforms often allow degrees of anonymity that make such a task impossible to do with any reliability.

Importantly, we cannot rely on social identity alone to determine whether a person will be offended by language or not. Social identity groups are not monolithic and identity likely has varying ability to predict dynamics in different geographic regions or for members new to group sociality. In addition, social identity is fluid across contexts and time. An obvious example is a change in one's age over time, but even at the same age, one may identify as young or old relative to the individuals they are with. It may make sense to consider social identity in relation to the specific temporal context of an utterance, yet headline news stories about prominent individuals who have lied about their social identities, such as Rachel Dolezal[3], offer a clear example of cases where social interactions once considered innocuous undergo re-evaluation. In other words, what is the identity that should be annotated or documented, and what bearing should this have on future classification of this language? More broadly, there are hard limits to inferring cultural context on the global web, which introduces challenges to identifying potential harms of offensive language, and adds difficulty to observing or measuring these ends compared with those of individual harms. For these reasons, user interface components that allow users to provide direct input on potentially offensive content remain valuable.

Detection and moderation practices that are unable to distinguish sociological patterns underpinning mock impoliteness stand to target it as well as underlying practices of reclaiming language. Indeed, Oliva et al. (2020) precisely raise censorship of drag queens' mock impoliteness as motivation for their work. Evidence of racial biases in offensive language classification and reports of the negative impacts of 'race-blind' approaches to content moderation also suggest that poorly targeted detection approaches may disproportionately impact marginalized communities. Due to the limits introduced by detecting social identity and observing platform behavior, language models are unlikely to identify mock impoliteness language in all cases. In practice, selecting a set of annotators aligned with the communities and sociological norms represented in data is nontrivial. There are also limits to the types and amounts of sociodemographic information that can or should be collected about data annotators and users. However, as offensive language classification improves, insights into providing context in annotation can also serve to shape content moderation processes, for example, by incorporating similar context for human review of content. In this respect, our work contributes to the development of frameworks of analysis attuned to sociological and historical modes of discourse that are critical for the responsible deployment of offensive language classification tasks.

[3] https://www.theguardian.com/us-news/2017/feb/25/rachel-dolezal-not-going-stoop-apologise-grovel

5 Conclusion

Contrary to common understandings of offensive language as negative and harmful, we show that offensive speech can function as a practice of resistance to unjust social norms and, in specific cases, can serve a socially beneficial role. In doing so we highlight three necessary criteria for evaluating offensive language, 1) the subject of an offensive utterance and their social position, 2) the outcomes of offensive language, and 3) the sociological role that offensiveness and offense serves. Queer mock impoliteness specifically illustrates that, although sarcasm, humor, and irony pose significant challenges to existing classification tasks, there is an ethical and social need to account for subversive uses of denotatively offensive language. This type of reappropriated speech serves to solidify a collective identity, protect ingroup members from outgroup abuse, and resist exclusionary and restrictive social norms. The practice finds its historical emergence in a different yet related practice of training for adversity put forward by the Cynics in which offensive discourse works as a way to challenge unreflective societal norms. Although operationalizing a relational definition of offensiveness comes with challenges, such as practical and ethical limits to observing social identity and user behavior, we point to promising research directions to better account for the expressly beneficial sociological role that offensiveness can play in social discourse.

References

Hannah Arendt. 2013. *The human condition*. University of Chicago press.

A Ashwitha, G Shruthi, HR Shruthi, Makarand Upadhyaya, Abhra Pratip Ray, and TC Manjunath. 2021. Sarcasm detection in natural language processing. *Materials Today: Proceedings*, 37:3324–3331.

Agathe Balayn, Jie Yang, Zoltan Szlavik, and Alessandro Bozzon. 2021. Automatic Identification of Harmful, Aggressive, Abusive, and Offensive Language on the Web: A Survey of Technical Biases Informed by Psychology Literature. *ACM Transactions on Social Computing*, 4(3):1–56.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2):100205.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1217–1230.

Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. 2014. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of communication*, 64(2):317–332.

Patricia Hill Collins. 2002. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. routledge.

A Feder Cooper, Benjamin Laufer, Emanuel Moss, and Helen Nissenbaum. 2022. Accountability in an algorithmic society: Relationality, responsibility, and robustness in machine learning. *arXiv preprint arXiv:2202.05338*.

Jonathan Culpeper. 2011. *Impoliteness: Using language to cause offence*, volume 28. Cambridge University Press.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2021. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *arXiv preprint arXiv:2110.05719*.

Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan K. Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. Association for Computing Machinery.

Michel Foucault. 1982. The subject and power. *Critical inquiry*, 8(4):777–795.

Michel Foucault. 2012. *Discipline and punish: The birth of the prison*. Vintage.

Michel Foucault and Michel Foucault. 2012. *The courage of the truth: (the government of self and others II). Lectures at the collège de France, 1983 - 1984*. Palgrave Macmillan, Basingstoke.

Danielle Gaucher, Brianna Hunt, and Lisa Sinclair. 2015. Can pejorative terms ever lead to positive social consequences? The case of SlutWalk. *Language Sciences*, 52:121–130.

Pierre Hadot. 2002. *What is ancient philosophy?* Harvard University Press.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 752–762.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)*, pages 483–488.

Dirk Hovy and Diyi Yang. 2021. The Importance of Modeling Social Factors of Language: Theory and Practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Alice E Marwick and Danah Boyd. 2011. I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society*, 13(1):114–133.

Sean McKinnon. 2017. "Building a thick skin for each other": The use of 'reading'as an interactional practice of mock impoliteness in drag queen backstage talk. *Journal of Language and Sexuality*, 6(1):90–127.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Author profiling for hate speech detection. *arXiv preprint arXiv:1902.06734*.

Stephen O Murray. 1979. The art of gay insulting. *Anthropological Linguistics*, 21(5):211–223.

Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & Culture*, 25(2):700–732.

Desmond Patton, Philipp Blandfort, William Frey, Michael Gaskell, and Svebor Karaman. 2019. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity Detection: Does Context Really Matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On Releasing Annotator-Level Labels and Information in Datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Farzana Rashid, Tommaso Fornaciari, Dirk Hovy, Eduardo Blanco, and Fernando Vega-Redondo. 2020. Helpful or hierarchical? predicting the communicative strategies of chat participants, and their impact on success. In *EMNLP| Findings*. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. 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, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. *arXiv preprint arXiv:2111.07997*.

Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–35.

Tones Smith. 2012. Pathology, bias and queer diagnosis: a crip queer consciousness.

Yujian Sun, Ying Li, and Tingxuan Zhao. 2021. The improved neural network model in humor detection with traditional humor theory. In *2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*, pages 549–554. IEEE.

Salome Viljoen. 2021. A relational theory of data governance. *Yale LJ*, 131:573.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Max Weber. 2019. Economy and society. In *Economy and society*. Harvard University Press.

Leila Weitzel, Ronaldo Cristiano Prati, and Raul Freire Aguiar. 2016. The comprehension of figurative language: What is the influence of irony and sarcasm on NLP techniques? In *Sentiment analysis and ontology engineering*, pages 49–74. Springer.