Relationality and Offensive Speech: A Research Agenda

Razvan Amironesei*
Unaffiliated
amironesei@gmail.com

Mark Díaz*
Google Research
markdiaz@google

Abstract

We draw from the framework of relationality as a pathway for modeling social relations to address gaps in text classification, generally, and offensive language classification, specifically. We use minoritized language, such as queer speech, to motivate a need for understanding and modeling social relations– both among individuals and among their social communities. We then point to socio-ethical style as a research area for inferring and measuring social relations as well as propose additional questions to structure future research on operationalizing social context.

1 Introduction

In this paper, we build on NLP-based approaches to defining and classifying offensive speech to lay out research directions for robustly incorporating social context into the ways text classification tasks are conceptualized and operationalized. Our motivation lies in classifying sociolinguistic norms of minoritized communities, such as the use of reclaimed slurs, which current classification approaches often fail to distinguish from language which is abusive, toxic, or hateful. To achieve a robust understanding of social context, we consider offensive speech in terms of relationality– or the social relations that inform how language is used and interpreted. At a conceptual level we defined offensiveness as a property of social relations rather than as a property of specific language terms. At an operational level, we discuss initial insights and open research directions for how social relations can be measured in practice.

Reclaimed language use and other aspects of minoritized language, such as queer speech and Black American vernacular have proven challenging for text classification (Dias Oliva et al., 2021; Sap et al., 2019). This language use reflects a plurality of language meaning and non-normative use that many NLP approaches currently fail to capture. The research directions we propose are oriented toward text classification for potentially harmful or undesirable speech, such as toxicity detection or hate speech detection. While we consider offensive speech to be distinct from hate speech or toxic language, they have important similarities that help to clarify a definition of offensive speech as well as point to approaches for improving classification tasks (Diaz et al., 2022). That is to say, while we use a definition of offensive speech that overlaps with definitions of hate speech and other abusive language, a sociological understanding of offensive speech indicates that it is distinct in ways that current classification approaches do not reflect. Our overarching goal is to provide research directions toward contextually-informed modeling and annotation to appropriately capture sociolinguistic norms used within minoritized groups. A key underlying postulate of our research is that speech, and in particular offensive speech, is not divorced from "doing". On the contrary, offensive speech has practical effects that enact and perform subjective formations (Butler, 2021).

Although a range of definitions and labels have been used to operationalize offensive language, they share a goal of classifying undesirable language that stands to harm or deteriorate discourse. Concepts for classification have included, “abusive language” (Nobata et al., 2016), “harmful speech” (Faris et al., 2016), and “hate speech” (Schmidt and Wiegand, 2017), among others. These tasks do not use identical definitions of offensiveness but often use similar labels and share similar goals. Definitional differences can be observed in the label schema for each task. For example, Van Hee et al. (2015) define ‘racist’ and ‘sexist’ as subsets of ‘insults’, and Wulczyn et al. (2017) include a specific label for personal attacks.

Importantly, researchers have identified issues and challenges related to the variety of social

*Authors contributed equally

The 7th Workshop on Online Abuse and Harms (WOAH), pages 85–95

contexts in which classification tools are applied, namely those involving satire and nonstandard use of language, such as reclaimed speech (Davidson et al., 2017). These challenges are rooted in nuanced use and understanding of language that rely heavily on aspects of social context including, culture, place, and power. Additionally, they point to a need for better incorporation of social context in the ways that NLP tasks are conceptualized and operationalized (Hovy and Yang, 2021a).

We refine this line of research by emphasizing that a socio-ethical account of offensive speech should be attentive to a diversity of contextual uses and the variety of forms it can take. This requires a basic understanding that the offensiveness of speech is dependent upon 1) the background of social and cultural conditions that surround it; 2) the social dynamics between the subjects and objects of offense; 3) the in-group/out-group language norms surrounding language use; and 4) the different types of outcomes of offensive speech, including the resulting potential and actual harms associated with the previous considerations. Our approach expands from (Diaz et al., 2022), who use conceptual analysis to evaluate specific components of how hate speech and toxicity are defined in order to form the basis for an expanded definition of offensive speech. Rather than an exhaustive review of definitions, they identify those which help to build a more robust approach to defining offensiveness, with specific attention toward identifying and operationalizing its relational qualities. Building from their work, we propose relationality as a conceptual bridge to more robustly operationalize social context and, in particular, the social relations that differentiate minoritized speech from antagonistic forms of speech. In addition, we point to existing work on style measurement as an avenue to do so.

In the following sections we draw from the framework of relationality to motivate a need for modeling social relations to address gaps in text classification, generally, and offensive language classification, specifically. Second, we propose research domains and questions to structure future research on operationalizing social context. Third, we point to and discuss examples for how we can begin to better model social relations. We do not provide a closed or exhaustive set of techniques for applying a relational lens, however we discuss style and its use in NLP as a jumping off point for addressing ethical concerns surrounding offensive language classification that others have raised (e.g., (Dias Oliva et al., 2021; Diaz et al., 2022).

2 A Relational Framework for Contextual Analysis of Offensive Speech

Relationality operates as a general analytic tool that helps to unveil and disambiguate specific contextual uses of offensive speech from others. A relational lens in the context of NLP refers to a focus on the social relations that influence the production, meaning, reception, and outcomes of language among interlocutors. In this way, relationality is a means of analysis to conceptually organize social context. Hovy and Yang (2021) propose to shift NLP analysis toward a contextual understanding of speech that consists in the following seven factors: 1) speaker and 2) receiver, 3) social relations, 4) context, 5) social norms, 6) culture and ideology, and 7) communicative goals (Hovy and Yang, 2021a). We argue that contextual analysis of offensive speech can be achieved through a focus on the social relations inherent in language, its use, and its outcomes.

Diaz et al. (2022) point out a distinction between treating offensiveness as a property of an utterance rather than as a relation between individuals or communities and that utterance. Treating offensiveness as a property of a linguistic token, such as by registering a term to a blocklist, ignores the very real ways in which language meaning is not fixed or inherent to its orthography but rather is constructed socially via a network of meanings among social actors. For this reason, when we refer to "offensive speech" we refer not only to the content of an utterance, but also the confluence of social relations and context that surround the production of that utterance. In other words, "offensive speech" entails time, place, by whom, and to whom, in addition to orthography. Relationality also reflects a move away from locating offensiveness exclusively at the level of words and instead locates offensiveness in an individual or group's relation to a word or concept. This, in turn, helps to distinguish why a term might produce offense when used between members of different communities but not when used between members of the same community, as in the example of reclaimed slurs.

Through relationality our focus is on accounting for patterns inherent in the social relations that pro-

duce offensive speech. In this respect, our work overlaps with prior work that effectively operationalizes aspects of relationality through analyses of interactional patterns and discourse (e.g., (Danescu-Niculescu-Mizil et al., 2011)). Relationality itself does not provide a comprehensive list of all the contextual elements that influence how communication is understood between social actors, however, it emphasizes how to conceptually organize social context– namely around social relations between and surrounding subjects. As such, applying relationality rests on further research and validation of the relevant aspects of social relations that must be accounted for across text classification tasks.

While Hovy and Yang (2021b) have laid important groundwork for addressing this question and Diaz et al. (2022) explore social context more specifically in the context of offensive language classification tasks, we propose several research directions for bringing relationality into practice for classification tasks. There has not been explicit work on detailing the aspects of social context most operative for distinguishing the range and differential impacts of offensive language. Each of these directions has overlapping components but address open questions about what a relational lens means for 1) how offensive language can be conceptualized in a way that is responsive to minoritized speech and 2) how offensive language is operationalized through annotation task design and language modeling.

2.1 Minoritized Speech

A problem we draw from that exemplifies the need for a relational lens is that posed by minoritized speech, which classification systems have been shown to misclassify or classify in systematically biased ways. For example, scholars in NLP have high error rates for African American English (AAE) in part-of-speech tagging and language identification (Jørgensen et al., 2015; Blodgett et al., 2016), and disproportionately toxic ratings of speech containing features of AAE compared with speech that does not (Sap et al., 2021). Another example is that of drag queen speech, which Dias Oliva et al. showed was rated more likely to be 'toxic' compared with tweets from white supremacists in a comparative study (Dias Oliva et al., 2021). As Dias Oliva et al. (2021). discuss, the discourse used by drag queens on Twitter is expressed through shared slang, references, and linguistic norms. Diaz et al. (2022) point out that using this language relies on shared assumptions about the use of slurs, mutual consent to break normative rules of language "decency", and an understanding that manners of speaking used in an in-group context can be qualitatively distinct from the use of those manners of speaking in an out-group context.

We understand minoritized speech as a type of speech that emerges as a result of a power asymmetry that is produced by dominant and widely accepted forms of expression within a language. Both Dias Oliva et al. (2021) and Diaz et al. (2022) note that communication in the queer community involves the reappropriation of offensive language as a means to "self-inoculate" community members against social attacks from out-group members. The same cannot be said about white supremacist speech which is defined by objectifying and demeaning historically marginalized groups and incitement of hate and violence (Blazak, 2009). The problem they raise, however, is not limited to the minoritization of drag queen speech. They argue that addressing the risk of increased censorship for minoritized language is an ethical imperative because of the socially productive role that non-normative language plays in the survival of minoritized groups (Diaz et al., 2022).

3 Relationality through Style

In response to the challenges posed by minoritized speech, we turn to linguistic style and its measurement in NLP as a means of both describing and applying relationality. In doing so, we draw from style as an artifact of social context that specifies how social relations are structured. Work on linguistic style in NLP has typically focused on individual communication style, such as in investigating author attribution (Safin and Ogaltsov, 2018) or making inferences about author psychological state and demographics (Pennebaker, 2011). Notably, measurements of style are usually pursued in contrast to explorations of language content. Khalid and Srinivasan (2020) bridge the gap between structure and content by applying style measurement to understand an individual's relationship to a broader community. The authors use style to explicate a social relation that is not necessarily explicit in an utterance itself. This moves from simply applying style to characterize individuals to understanding a

broader social relation and orientation to community language norms.

Our contention is that NLP accounts of style must explicitly contend with the social, historical, and practical conditions from which styles of speech emerge. Thus, work on style in NLP needs to be attuned to the underlying ethical questions associated with the technical measurement of styles of speech. First and foremost, this means that style needs to be understood as embedded in specific contexts of production with distinct practical outcomes. For example, at an ethical level, style can be understood: (1) as the reflexive practice of styles of existence via the exercise of specific technologies of the self (Martin et al., 1988; Hadot, 1995), such as practices of self writing (Foucault, 2019) and practices of truthful speech (Foucault, 2011); (2) as a work of forming and transforming one's existence (Foucault, 2012; McWhorter, 1999) via somatoaesthetic projects that are not reducible to the purely individual and voluntaristic manifestations of heroic self-distinction (Heyes, 2007) and moral quests for universal wisdom predicated on self-possession (Amironesei, 2014). However, an ethical grounding of style is related yet distinct from a strictly sociological (Fleck, 2012; Zittel, 2012), historical (Crombie, 1994) and an epistemological account of styles of thinking (Hacking, 1992). From an ethical standpoint style is conceptualized as a practice of the self and others while at a sociological level, style is the product of community language norms that reflect hierarchical patterns of discourse that are interwoven with social identity formation and relational dynamics (Labov, 1973). In both cases, a socio-ethical account of style is context-dependent, "relational and dynamic" (Ekström et al., 2018) and a key feature of an individual or a group's self-expression. One aspect that we emphasize here is that style has irreducible ethical, social and political conditions, expressions and manifestations which refer to speech that an individual or a group produce in relation to others, rather than as a fixed property of an individual, their words or given images. In this way, analyses of style can be robust to code-switching or the range of styles individuals may use in changing social contexts.

Thus, given the contingent and contextual production of style we propose relationality through style as an analytic or a mode of analysis that seeks to account for the historically and socially constituted matrices of power relations where style works as an interactive feature which opens to spaces of contestation in the formation of both individuals and collectives. For our purposes, while style provides general indications of social context, its relational significance lies in its potential for disentangling minoritized forms of speech from abusive language. For example, mock impoliteness, which features in drag queen speech, plays a central role in group identity formation and resistance against oppressive social systems (McKinnon, 2017). Style's significance for minoritized communities emerges through "contextualized repertoires of speaking and behaving through which identities and sociocultural affiliations are claimed and communicated" (Ekström et al., 2018).

A key takeaway is that a common style among interlocutors can suggest shared norms or social or cultural proximity. Because style is an artifact of social norms, and thus social relations, it can be used to infer shared context among individuals involved in an interaction being assessed for offensive content. In this way, comparisons of linguistic form can be a tool to unveil the relations among which offensive language is couched. While style can vary from individual to individual, Khalid and Srinivasan (2020) show that style can reliably predict group membership, independent of language content (Khalid and Srinivasan, 2020). Indeed, earlier work has shown that style can indicate social demographic information about a speaker (Eckert and Rickford, 2001). In the context of offensive speech classification, this means that style provides useful information for assessing whether individuals share a sociolect or dialect. This carries significance not only for disambiguating language use within a given minoritized sociolect and improving upon weaknesses in offensive speech classification.

3.1 Style and Common Sense

By failing to disambiguate language uses, particularly those that are minoritized, current classification approaches implicitly force a generalized or 'common sense' interpretation of language, whether at train time (i.e., via annotation) or at inference time. Using style measurement, or other relational approaches, to situate language explicitly in its social relations puts into practice the understanding that the same language can carry different connotations or meanings. A pluralistic understanding of language is not possible through approaches that ignore relations between individuals and the

communities they belong to. This is because, in the absence of explicit, familiar sociolinguistic cues or relational context, a reader or model must interpret the language using generalized language norms as a primary point of reference thus concealing the differential relations (Deleuze, 1994; Boven, 2014) that occur between various ways that individuals and communities engage with language.

From a ML fairness perspective, applying generalized language norms is to rely on dominant, prescriptivist views that often treat minoritized speech as improper and contribute to biased system performance. One example of stigmatized speech, AAE, has been characterized as incorrect, devaluing not only its use, but also the communities that speak it. As Pullum demonstrates, AAE "is not Standard English with mistakes," rather its "speakers use a different grammar clearly and sharply distinguished from Standard English at a number of points" (Pullum, 1999). In NLP, Sap et al. (2019) show that "AAE tweets are twice as likely to be labeled offensive compared to others" and recommend paying special attention to the effect of a speaker's dialect and social identity to mitigate negative and disparate impacts. Aside from being ethically dubious, applying generalized language norms drawn from prescriptivist views of language use ignores nuances and distinctions between uses of AAE in in-group and out-group contexts.

We eschew any analysis that treats language as offensive based on guidelines grounded in notions of common sense or, in the case of offensive language classification, notions of common decency or civility. This is precisely because notions of common decency, like notions of generalized language norms, stand to devalue minoritized sociolinguistic norms. Civility is not always explicitly defined in text classification contexts, but has been articulated as "concerned with communicating attitudes of respect, tolerance and consideration to one another' (Calhoun, 2000). While common sense can indicate general, accepted uses of speech, it is culturally and contextually dependent, and thus falls within the set of factors that a relational lens is needed to disambiguate, including the subjects and relations that they are embedded in. Without disambiguation, applying generalized language norms stands to be exclusionary by reflecting stigmatizing beliefs about non-standard language. At the same time, notions of common sense are vague and difficult to define as well as ignore the variety of conditions and contexts in which language is used.

The conceptual distinction between a relational and a common sense approach to language processing is not a mere abstraction that NLP researchers should simply be aware of. On the contrary, it has major implications for language modeling processes. For example, annotating the presence of offensive language in a rating task, with limited social context posits that there is a widely understood corpus of offensive language that a rater can draw upon that is distinct from another corpus of non-offensive language that represents decent, and civil discourse. The problem with this distinction is that offensive speech is historically constituted, that is, offensive terms change over time, and are defined by societal and cultural norms and power relations between groups. Annotators may draw from overlapping notions of civil language, however a variety of speech exists outside of these norms.

The measurement of style to study an individual's relationship to a broader community and its communication norms in the way that Khalid and Srinivasan (2020) do provides motivation for measuring the relationships among speakers across different communities. While style does not necessarily speak to specific relationships between individuals, overlaps in style can suggest some degree of shared norms or values. Still further research is needed to better understand how style might be used alongside other information collected or inferred in NLP tasks. For example, in their study of bias in toxicity ratings for AAE, Sap et al. (2019) showed raters an estimation of a tweet's likelihood to contain elements of AAE as well as primed raters to consider dialect in relation to the author's likely racial identity. They found that raters provided less biased toxicity annotations of AAE tweets after their intervention. It is not known whether the score caused raters to re-interpret the text examples according to AAE norms or if raters adjusted their annotations out of fear of appearing racist. However questions remain about why exactly the intervention succeeded and whether rater subgroups were similarly impacted by the intervention. Determining how relational approaches can best be applied to operationalize social context raises a number of research directions that we outline in the following sections.

4 Operationalizing Relational Context

In practice, the primary challenge of applying a relational approach to offensive language lies in defining its scope and operationalizing its component parts. As others have pointed out, identifying social context and integrating it into NLP models is both needed for more robust and successful NLP as well as nontrivial (Hovy and Yang, 2021a). Measuring linguistic style provides one way of applying a relational lens, however other features may be leveraged to infer relational context.

A relational focus in classification tasks requires determining a set of measurable features that provide information about social relations, as well as work to prioritize features that most improve task performance, particularly with respect to language norms missed by current techniques. Identifying and predicting aspects of social context as a part of classifying offensive speech brings its own, deep set of research questions and challenges. Although offensive speech detection and related tasks have largely been framed as text classification tasks, we break down research questions for future work into those that focus on linguistic features and those that focus on extra-linguistic features surrounding text and its production. In doing so, we implicitly shift offensive language classification from a text classification task to one that expands to include non-textual inferences in addition to linguistic content. Through a relational lens, language is one artifact produced by the social relation of offense between social actors. Framing language in this way allows us to consider other artifacts that result or shift as a consequence of offense. This broadens the range of features at our disposal to infer social context, including user behavior (e.g., "liking" comments), networks of user accounts, the post structure of dialogue, and histories of interactions. Taking advantage of this broadened set of features, we propose areas for research that build both from established approaches in language modeling, such as text annotation, as well as modeling approaches focused on non-text data, such as conversation structure, can serve as a clue to the nature of the relationship between two social actors (Zhang et al., 2018).

4.1 Context through Linguistic Features

As previously described, existing NLP techniques for modeling linguistic style and language dialect implicitly carry information about cultural context and community membership and should be further explored for the relational insights they bring. However, prior challenges in text classification, such as classifying reclaimed speech or satire, also bring to light research opportunities with respect to capturing social context at the data annotation step. Though not exhaustive of all opportunities for improving capture of social context, text annotation and annotation task design are ripe for additional work. Further research in these areas will be key to operationalizing relational aspects of language, precisely because human annotation is well-suited for capturing explicit social dynamics and interpretations that automated methods struggle with.

4.2 Context through Annotation

We bring a focus to annotation because the complexity of social context provides an opportunity to leverage human inference. Annotation tasks are typically designed in such a way that they isolate examples from the social context in which they were produced. This modularity makes the annotation of large volumes of data more efficient, but also introduces difficulties for data annotators who may lack important context in order to select an appropriate label for a given example. This also effectively takes a problematic common sense approach.

With respect to queer vernacular and erroneous classifications of toxicity, one reason for these misclassifications likely lies in idiosyncratic uses of otherwise offensive language in queer vernacular, such as the use "b*tch" or "f*ggot" as consensual terms of endearment. Idiosyncratic uses of language, including reclaimed speech, raise questions about how this language use can be made apparent to workers and distinguished from language use in other sociolects. As McKinnon notes, failing to distinguish this language brings with it ethical issues rooted in the fact that this language constitutes a means of queer survival (McKinnon, 2017). As a first direction of research focused on data annotation we ask: **How can additional context be provided in annotation tasks to support raters in understanding the original relations surrounding text examples?** Moreover, **what influence does additional information have on both annotation behavior and model performance?**

Some researchers have experimented with reintroducing social context into annotation tasks with varying degrees of success, such as by provid-

ing multiple turns in a conversation or exchange (Gao and Huang, 2017; Sap et al., 2019; Pavlopoulos et al., 2020). This stands in contrast to typical annotation approaches, which require raters to judge whether an utterance is offensive without context apart from what is contained within it. Utterances may often name the target and receiver, and can offer some cultural, demographic, and ideological context if it is named explicitly; however the social relations are particularly difficult to infer from an isolated message. There are opportunities to experiment with other kinds of social context, such as the website or origins of text examples and temporal information about when the interaction occurred. At the same time, it is important to explore the limits to what kinds of social context can be provided to raters, whether due to knowability or privacy preserving limitations.

Thus far investigations of providing social context in annotation tasks have taken a quantitative focus to measure if and how additional context changed the resulting annotations collected. Simply providing raters with more context may be of little value if the raters themselves lack social or cultural awareness of specific domains, such as queer life and vernacular. Thus, it is unclear how annotators use additional context when it is provided, the role their own social experiences play in their ability to understand sociolects or cultural references, as well as which kinds of examples require additional context for annotators to make confident assessments. Thus, as a complementary annotation research direction to the first we ask: **How do annotators understand and use contextual cues provided to them?**

This research direction builds, in part, from ongoing work in NLP and ML considering annotator diversity, social identity and their influence on the annotations raters provide (Díaz et al., 2022; Prabhakaran et al., 2021; Davani et al., 2022), as well as work that qualitatively investigates annotation work practices. For example, scholars such as (Miceli et al., 2020) have provided rich accounts of how organizational structures influence how annotators are able to conduct their work. As a result, researchers seeking to understand annotation behavior must consider factors beyond what they can measure through typical metrics such as inter-rater reliability. This work is undertaken with a motivation to understand variation in annotation behavior and its potential roots in the social experiences, sociodemographics, and labor contexts of workers. In the context of queer vernacular, it would be intuitive to expect that a queer rater is more likely to understand the social context of a text example involving queer speech compared with a non-queer rater, provided they are given sufficient context to begin with. (Díaz et al., 2022; Prabhakaran et al., 2021) make calls for reporting transparency for crowdsourced data collection so that dataset users can investigate systematic disagreements and representation.

However, there remain open questions regarding how raters might apply their 'interpretive lens' and, more broadly, how these perspectives might be incorporated reliably into data collection efforts given the sensitivity of questions regarding membership to minoritized communities. We propose this direction with a specific eye toward research that incorporates qualitative approaches and understandings of annotation work. Relational considerations regarding data annotation include not only the relations embedded in data examples, but also the social relations between annotators and content embedded in the data they annotate.

4.3 Context through Extra-Linguistic Features

In addition to research on annotation task design, we propose expanded exploration of modeling techniques. A relational approach on offensive language brings into focus not just the specific language used in an interaction, but also behaviors and context that surround an offensive interaction. These include the behaviors, such as an individual's past interactions with content (e.g., 'liking' or downvoting) and other users, and metadata that captures temporal and geocultural situatedness. Using these features and techniques as windows into social context, there are opportunities to additionally model extra-linguistic features to more robustly infer social context.

Features apart from those specifically embedded in the text of an utterance can be used to provide clues into relevant social context in an interaction. (Mishra et al., 2019) do precisely this in incorporating author profiles in their modeling of racist and sexist tweets. From author profiles, they were able to model user-specific information, such as their network of followers. This approach effectively ties a given utterance to be classified to the particular individual who produced it. This stands apart

from approaches which implicitly assume that an utterance carries the same offensive nature independent of who produced it. Mishra et al. specifically call out this shortcoming, arguing that deviations from sociolinguistic norms within communities is important for understanding the varied forms that abusive language can take (Mishra et al., 2021). Building from this work we ask:

What additional features become useful for identifying offensive content in the shift from targeting offensive speech to targeting offensive relations? and **How might extra-linguistic features, such as conversation structure and non-textual content, be used to infer social context?**

4.3.1 Interaction Outcomes

With the notable exception of work on toxicity, little work has focused on measurable outcomes of offensive interactions. Toxicity is a prediction of whether a tweet or excerpt of text will cause its audience to disengage from an interaction. This provides a proxy for determining the inappropriate or offensive nature of text that can be measured through behavior. (Diaz et al., 2022) point out that maintaining user engagement may not be desirable and may have disproportionately negative consequences for minoritized users. We focus, however, on the incorporation of non-textual observations that toxicity inspires. These include outcomes of interactions, such as downvotes and blocking user profiles, as well as behaviors that precede a given interaction, such as a user's past posting behavior. Additionally, (Zhang et al., 2018) use conversation structure in modeling whether user interactions on Facebook Pages will result in users blocking one another. Using an extended conversation as a unit of analysis opens up opportunities for modeling interactions and additional social context. As a complement to work on how text should be annotated we propose another research direction that asks: **What are relevant interaction outcomes that can be measured and used to model interactions that produce offense?**

(Mishra et al., 2019)'s work points to additional opportunities to assess individuals' communication history in relation to one another. For example, patterns in individuals' communication history may indicate repeated, antagonistic behavior. A related area of work lies in online trolling detection, which has been pursued through user-based methods, post-based methods, thread-based methods and social network analysis (Tomaiuolo et al., 2020). While not all offensive language falls under the umbrella of trolling, techniques used to detect trolling highlights avenues for measuring behaviors in relation to offensive language use.

5 Conclusion

Our chief claim is that relationality and its sociological and ethical formulations of linguistic style are a promising guiding analytic for achieving a more robust contextual analysis of offensive speech. Motivated by the challenges posed by minoritized language norms, we propose avenues for research that take aim at operationalizing it in practice. Because style patterns can be used to unveil social relations among individuals and communities, we point to its measurement as an example for operationalizing our approach. Ultimately, offense is produced through social relations that must be ethically and sociologically understood in order to accurately model and classify language content. Focusing on social relations and their potential to help distinguish sociolinguistic norms generates the following research questions:

- What are the relevant aspects of social relations that must be accounted for across text classification tasks?
- How might structural elements of style, which have been measured in various ways, be complemented by measurements of sociological and ethical aspects of style?

In the context of data annotation:

- How can additional context be provided in annotation tasks to support raters in understanding the relations surrounding text examples?
- Moreover, what influence does additional information have on both annotation behavior and how do annotators use contextual cues provided to them?

With respect to modeling language and social interactions:

- What additional features become useful for identifying offensive content in the shift from targeting offensive speech to targeting offensive relations?
- How might extra-linguistic features, such as conversation structure and non-textual content, be used to infer social context?

- What are relevant interaction outcomes that can be measured and used to model interactions that produce offense?

Importantly, we must also explore the limitations of a relational approach rooted in style. While style has important connections to the formation of individual and collective identities, it has different uses, such as to comply with institutional (e.g., workplace) norms, which may not necessarily align with community norms or social identification processes. In addition, style can be deployed in adversarial ways, such as with mockery, intent to impersonate, exploit trust, or arguably 'inauthentic' uses of accent or dialect (e.g., appropriative use of a "blaccent" (Lockhart, 2021)). It is unclear, at least at an operational level, how relationality might account for these uses of style. Another complication lies in the fact that in many digital contexts, one's "true" identity is often not verifiable. For our purposes, this means that a person can communicate online using styles of speech that may align with offline specific manners of speaking (e.g., AAE) but that may not align with the styles they use in other contexts. Underlining all of the above limitations is a greater tension regarding the ethical risks of inferring social identity and the extent to which inferring an individual's social identity is meaningful for classification. Hamidi et al. (2018) studied trans* and gender nonconforming individuals' perceptions of automatic gender recognition systems, demonstrating how automated systems can contribute to misgendering harms and undermine individual autonomy. Thus, inferences about social context that rely on further inferences about social identity

Still, relationality can work as a frame of analysis for the design of NLP approaches, including annotation practices and modeling decisions that can unveil specific relational context. We have identified minoritized speech as a motivating example to show how current, generalized approaches are inadequate for classifying language that deviates from dominant sociolinguistic norms. Providing sound criteria to disambiguate and classify a plurality of modes of speech grounded in a deep social understanding of their formation is key to ensure a more just and ethical approach to offensive speech.

References

Razvan Amironesei. 2014. La déprise de soi comme pratique de désubjectivation: Sur la notion de "stultitia" chez michel foucault. *Journal of French and Francophone Philosophy*, 22(2):104–122.

Randy Blazak. 2009. Toward a working definition of hate groups. *Hate crimes*, 3(1):133–162.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*.

Martijn Boven. 2014. 11 a system of heterogenesis: Deleuze on plurality. In *Phenomenological Perspectives on Plurality*, pages 175–194. Brill.

Judith Butler. 2021. *Excitable speech: A politics of the performative*. routledge.

Cheshire Calhoun. 2000. The virtue of civility. *Philosophy & public affairs*, 29(3):251–275.

Alistair Cameron Crombie. 1994. *Styles of scientific thinking in the European tradition: The history of argument and explanation especially in the mathematical and biomedical sciences and arts*, volume 2. Duckworth.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words! linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Gilles Deleuze. 1994. *Difference and repetition*. Columbia University Press.

Thiago Dias Oliva, Marcelo Antonialli Dennys, 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.

Mark Diaz, Razvan Amironesei, Laura Weidinger, and Iason Gabriel. 2022. Accounting for offensive speech as a practice of resistance. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 192–202, Seattle, Washington (Hybrid). Association for Computational Linguistics.

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.

Penelope Eckert and John R Rickford. 2001. *Style and sociolinguistic variation*. Cambridge University Press.

Mats Ekström, Marianna Patrona, and Joanna Thornborrow. 2018. Right-wing populism and the dynamics of style: a discourse-analytic perspective on mediated political performances. *Palgrave Communications*, 4(1):1–11.

Robert Faris, Amar Ashar, and Urs Gasser. 2016. Understanding Harmful Speech Online. *SSRN Electronic Journal*.

Ludwik Fleck. 2012. *Genesis and development of a scientific fact*. University of Chicago Press.

Michel Foucault. 2011. *The courage of truth*. Springer.

Michel Foucault. 2012. *The history of sexuality, vol. 2: The use of pleasure*. Vintage.

Michel Foucault. 2019. *Ethics: subjectivity and truth: essential works of Michel Foucault 1954-1984*. Penguin UK.

Lei Gao and Ruihong Huang. 2017. Detecting Online Hate Speech Using Context Aware Models. In *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, pages 260–266. Incoma Ltd. Shoumen, Bulgaria.

Ian Hacking. 1992. 'style'for historians and philosophers. *Studies in History and Philosophy of Science Part A*, 23(1):1–20.

Pierre Hadot. 1995. Philosophy as a way of life: Spiritual exercises from socrates to foucault.

Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–13.

Cressida J Heyes. 2007. *Self-transformations: Foucault, ethics, and normalized bodies*. Oxford University Press.

Dirk Hovy and Diyi Yang. 2021a. 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.

Dirk Hovy and Diyi Yang. 2021b. 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.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the workshop on noisy user-generated text*, pages 9–18.

Osama Khalid and Padmini Srinivasan. 2020. Style matters! investigating linguistic style in online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 360–369.

William Labov. 1973. *Sociolinguistic patterns*. 4. University of Pennsylvania press.

Amirah Lockhart. 2021. A stolen culture: The harmful effects of cultural appropriation.

Luther H Martin, Huck Gutman, and Patrick H Hutton. 1988. *Technologies of the self: A seminar with Michel Foucault*. Tavistock.

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.

Ladelle McWhorter. 1999. *Bodies and pleasures: Foucault and the politics of sexual normalization*. Indiana University Press.

Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–25.

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

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2021. Modeling users and online communities for abuse detection: A position on ethics and explainability. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3374–3385, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153, Montréal Québec Canada. International World Wide Web Conferences Steering Committee.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? *arXiv preprint arXiv:2006.00998*.

James W Pennebaker. 2011. The secret life of pronouns. *New Scientist*, 211(2828):42–45.

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.

Geoffrey K Pullum. 1999. African american vernacular english is not standard english with mistakes. *The workings of language: From prescriptions to perspectives*, pages 59–66.

Kamil Safin and Aleksandr Ogaltsov. 2018. Detecting a change of style using text statistics. *Working Notes of CLEF*.

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*.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Michele Tomaiuolo, Gianfranco Lombardo, Monica Mordonini, Stefano Cagnoni, and Agostino Poggi. 2020. A survey on troll detection. *Future internet*, 12(2):31.

Cynthia Van Hee, Ben Verhoeven, Els Lefever, Guy De Pauw, Véronique Hoste, and Walter Daelemans. 2015. Guidelines for the fine-grained analysis of cyberbullying.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399, Perth Australia. International World Wide Web Conferences Steering Committee.

Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations gone awry: Detecting early signs of conversational failure. *arXiv preprint arXiv:1805.05345*.

Claus Zittel. 2012. Ludwik fleck and the concept of style in the natural sciences. *Studies in East European Thought*, 64:53–79.