Constructing a Linguistic Classifier of Tone Policing

Adrian Ray-Avalani
Northwestern University
arayavalani@u.northwestern.edu

Abstract

Tone policing is a rhetorical strategy that involves using a speaker's tone or manner of expression as a reason to dismiss or invalidate the content of their message. Sociological work on tone policing generally agrees that it is an unconstructive conversational tactic that can cause harm both immediate and far-reaching, often linked to dynamics of power and oppression. Because tone policing is typically not explicitly hateful, it does not fall under the definitions of hateful or offensive speech detectable by many existing computational hate speech classifiers. However, despite this being a complex judgment task even for humans, our work shows that LLMs are able to relatively accurately identify instances of tone policing on Twitter, with >85% overall accuracy with all versions of the model tested. We also contribute to an increased linguistic understanding of tone policing by providing a typology of sentence constructions that typically characterize tone policing, and lay the groundwork for further investigation both of computational applications of tone policing detection or sociolinguistic research on who engages in tone policing and why.

1 Introduction

1.1 What is tone policing?

Nigerian-American writer Ijeoma Oluo, in her book "So You Want To Talk About Race" (2018), defines the rhetorical strategy of "tone policing" as follows:

> "Tone policing is when someone (usually the privileged person) in a conversation or situation about oppression shifts the focus of the conversation from the oppression being discussed to the way it is being discussed. Tone policing prioritizes the comfort of the privileged person in the situation over the oppression of the disadvantaged person."

Oluo's definition demonstrates both the unconstructive nature of tone policing in a conversation, as well as its harmful effects, especially when privilege and oppression are in play. However, by nature, tone policing is a relatively subtle form of potentially harmful speech. As Davidson et al. (2017) discusses, explicitly hateful speech such as the usage of slurs is easily identifiable as such by both humans and computational models, but tone policing may be couched in positive-seeming language. A person engaging in tone policing may even be doing so out of a genuine desire to help: the concept of respectability politics, for instance, refers to members of marginalized communities policing each others' manners of expression in order to hew to a third party's standard of "respectability." (Hill, 2018)

The field of natural language processing (NLP) has for some time worked on the problem of detecting hateful and offensive speech for the purpose of auto-moderation of online spaces. However, these models are largely neural-net based, resulting in model judgments which typically are neither human-comprehensible nor help researchers understand more about the usage of hate speech. Furthermore, Davidson and others note that categorization schemas of hateful and offensive speech vary significantly between models and even contradict one another; more subtle forms of harmful speech, like tone policing, are often not classified as harmful by such models. Furthermore, to our knowledge, no existing computational model of naturalistic harmful language includes tone policing as a possible label. Current understandings of tone policing as a concept have been developed in sociological literature as well as in popular culture discourse; however, tone policing has neither been studied from a linguistic perspective nor analyzed with computational tools.

1.2 Research Questions / Goals

Therefore, the overarching goal of the current work is to contribute an increased understanding of tone policing through a computational linguistic framework. In particular, we state the following guiding questions for our research:

RQ1: What linguistic structures characterize tone policing?

RQ2: Can a computational model accurately detect tone policing in text-based conversations?

RQ3: Are certain classes of linguistic structures more or less likely to be detectable as tone policing by a computational model?

Via a dataset of tweet-reply pairs from Twitter, we create a lexical rule-based classifier of tone policing from which we derive a human-understandable schema of sentence constructions that occur commonly in instances of tone policing. We subsequently finetune a DistilBERT-based computational model on our dataset, achieving good overall accuracy and improvements on baseline with all versions of the model.

In answering these research questions, our model contributes to the field of computational linguistics a more nuanced understanding of tone policing as a form of potentially harmful speech. Furthermore, our work demonstrates that it is possible for computational models to accurately detect whether a statement is engaging in tone policing or not, thereby providing a foundation on which tools for automatic detection of subtly harmful speech online can be built.

2 Related Work

2.1 Sociological

Prior academic work on tone policing has largely been conducted through a sociological lens, allowing access to quite a few naturalistic examples of tone policing as well as rich analyses of when and why it occurs, and the variety of harmful effects it may cause. Several papers examine tone policing as a type of racist backlash faced by women of color. Kwarteng et al. (2021) analyzes comments on Twitter directed at four Black women, all of whom had recently discussed their experiences on Twitter of being dismissed from major tech companies due to gender and/or racial discrimination. From their corpus of Twitter comments, Kwarteng et al. (2021) construct a model of misogynoir, i.e., the types of abusive or hateful speech that are directed at Black women due to a combination of racism and sexism. The authors define tone policing, one of their categories of misogynoir, as "a mechanism for preserving the status-quo through suppressing expressions of anger in response to injustice [...] One can identify Tone Policing when individuals critique the form and not the content of a serious message about injustice." Kwarteng et al. (2021)'s examples of tone policing comments include phrases such as "not constructive," "whining about," and "rude and arrogant way of speaking," used to describe the Black women themselves as well as their statements.

In her analysis of comments posted on Black-oriented blog or magazine articles, Davis (2020) points out a number of racist tropes including "'concern trolling,' which takes place when a user attempts to derail conversations about race and racism through insincere expressions of concern about the consequences of such discussions." Davis subsequently defines tone policing as a primary form of concern trolling, involving "the suggestion that discussions of racism result in more racism, with the implied exhortation that blacks stop talking about it." She provides an example of tone policing from the comments of an article titled "The Most Useless Types of White People, Ranked," from the Black culture blog VerySmartBrothas. A commenter writes:

> "Wow man, really helping your cause. I guess I will continue to be a "worthless white person. With shit like this racism will exist forever and I no longer care . . . if me trying to be decent makes me the butt of your joke, it's a great way to make me not try any longer."

This commenter is not literally criticizing the tone of the article, but is nevertheless engaging in behavior that polices tone or manner of expression: by insisting that their own hurt feelings be prioritized over the right of the article authors to discuss racism. This tendency is further elaborated upon in Nuru and Arendt (2019), who examine racial microaggressions directed from white women towards women of color. Nuru and Arendt (2019) emphasize that tone policing consists of "communication practices that prioritize the comfort of the privileged over the oppression of the disenfranchised," e.g. when "Leona [White participant] implies that

using a tone that is more palatable for White audiences is somehow better, and 'healthier,' than the way in which WOC [Women of Color] choose to communicate about their lived experiences."

The primary emotion objected to by Nuru and Arendt (2019)'s white participants is anger, which is variously construed as unproductive, as "hate," and as "put[ting] people on the defensive." This relationship between expressions of anger and their subsequent interpretations, especially as unproductive or potentially offensive, is also remarked upon in Biddle and Hufnagel (2019)'s case study of a very different environment.

Biddle and Hufnagel (2019) examine the case of answers to a survey on school values in a high school, and the discussions among student leadership groups about the data received. Early in the discussion, several student leaders voiced fears of teachers not taking the data seriously, or being pushed into the "danger zone," i.e. "what one enter[s] when pushed to an emotional place so uncomfortable that constructive dialogue or action could no longer happen." Several students described experiences of resistance from adult administrators that led them to feel "fear and anxiety about the messiness of managing adults' emotional reactions to any data on their teaching from students."

Because of this pervasive experience of student leadership efforts as "fragile and easily jeopardized by [teacher or administrator] resistance," the group therefore chose to tone police the survey responses, only highlighting positive or "politically correct" comments while setting aside ones seen as too "tough." Comments which strongly and bluntly expressed anger and other negative emotions about the school were deemed uncivil and therefore unproductive due to their proximity to the "danger zone." Biddle and Hufnagel (2019) conclude that the student leadership group "instituted a double standard for emotional expression within their work. They designated the bounds of civility to strategically and politically prioritize teachers' emotional well-being over sharing the students' concerns and emotional sense making."

Synthesizing the above sociological literature, we therefore construct our own working definition of tone policing for the purposes of this project. We define tone policing as an action in which the tone or manner of expression of a statement is used by a third party to justify invalidating the content of the statement, without actually engaging with said content. Our definition is as follows:

> Tone policing can be identified when:
>
> - a statement is made by one party, and
> - another party implicitly or explicitly criticizes the statement's tone or manner of expression, and
> - this party is implicitly or explicitly dismissive of the content of the statement, and
> - the dismissal occurs primarily due to tone or manner of expression, and
> - tone policing may also include:
>   - leveraging of a power dynamic between the parties
>   - combination with other types of speech, e.g. insults, humor/sarcasm, concern, positive framing
>   - some engagement with statement content - but crucially, dismissal is primarily not due to content, but due to tone

2.2 Computational

2.2.1 Hateful and Offensive Speech Detection

From our above definition of tone policing, we are therefore able to see that when a person engages in tone policing, they are not contributing to the conversation at hand. Rather than constructively engaging with the content of the conversation, they are insisting that their interlocutor's tone or manner of expression be modified (often to conform to dominant societal norms) in order for the content of the message to even be considered. Tone policing is therefore both unconstructive and actively detrimental to the original speaker's intentions in communicating their message; from the examples in the prior section, we see that it can even be explicitly offensive or toxic, especially when the party engaging in tone policing holds more societal privilege than their interlocutor.

Computational research on the problem of automatically detecting harmful and toxic speech online has been ongoing for some time. One such model comes from Mathew et al. (2022), whose dataset HateXplain relies on a classification of text into one of three categories: "hate speech," "offensive

but not hateful," or "normal (neither hateful nor offensive)." However, Mathew et al. (2022) also note that among the field of hate speech classifiers there is a large variety of categories into which speech is classified, e.g. "abusive," "hateful," "racist," "sexist," "toxicity," "spam," in addition to some datasets which use words like "abusive" and "offensive" interchangeably, or do not draw a distinction between hateful and offensive speech.

Mathew et al. (2022) derive their schema from Davidson et al. (2017), who define hate speech as "language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group [...] this may also be language that threatens or incites violence." They do not provide an explicit definition of offensive speech, but reference "people often [using] terms that are highly offensive to certain groups but in a qualitatively different manner." By contrast to this schema, Vidgen et al. (2021) present a dataset with just two labels, "hate" and "not hate." The authors acknowledge the limitations of only two categories; however, they also provide a detailed taxonomy of hate speech, as well as secondary categories for types of hate: "derogation," "animosity," "threatening language," "support for hateful entities," and "dehumanization." They also include a list of "high priority identities" that can be the target of hate speech (e.g. women, gender minorities, people with disabilities, refugees, immigrants). While Vidgen et al. (2021)'s taxonomy is quite thorough, this further emphasizes the problems in the field caused by so many differing, sometimes even contradictory definitions of harmful speech. One such problem is that subtler forms of harmful speech are often overlooked due to the lack of clear identification.

In Figures 1 and 2, sentences extracted from sociological literature on tone policing were submitted to the hate speech classifiers developed by Mathew et al. (2022) and Vidgen et al. (2021), respectively. The first sentence ("you have a rude and arrogant way of speaking") is derived from examples provided in Kwarteng et al. (2021). The second sentence ("I don't respect your statement because of the anger that came with it") is a quote from the singer Miley Cyrus directed at the rapper Nicki Minaj, which has widely been cited as tone policing in both academic and popular culture articles (e.g. Flores (2016); Johnson (2015); and others).

In Figure 1, the Mathew et al. (2022) model classifies both sentences as primarily "normal," with a smaller probability assigned to "offensive" and almost none to the category of "hate speech." In Figure 2, the Vidgen et al. (2021) model classifies both sentences as "not hate" with 100% certainty. These examples show how current hate or offensive speech detection models can often miss instances of tone policing, despite the demonstrable harmful effects caused by this type of speech when a power dynamic is present (as it is in both the example sentences used here).

2.2.2 Tone Policing Detection

Thus far, only two computational models appear to consider tone policing as a category of harmful speech. Parikh et al. (2019) seeks to classify accounts of sexism from the Everyday Sexism Project website, an online space where people who have experienced sexism can describe the incident(s) in their own words. Parikh et al. (2019) construct a taxonomy of sexism that includes the category of tone policing, which they define as "Comments or actions that cause or aggravate restrictions on how women communicate." Sheng et al. (2021)'s classifier of ad-hominem speech does not include tone policing as a subcategory of ad hominem on its own, but states that "We specifically include categories such as "ignorance" and "condescension" to cover more subtle forms of personal attacks (e.g., tone policing, mansplaining) that could further diminish the credibility of those who are already marginalized." Sheng et al. (2021)'s qualitative analysis of their naturalistic human dataset (tweets from Twitter) as well as of DialogGPT-generated data finds that human ad-hominem responses are most likely to fall in the categories of "condescension" and "ignorance," although their automated classifier does not label data with these subcategories, but simply as "ad-hominem" or "not ad-hominem." Aside from the work of Parikh et al. (2019) and Sheng et al. (2021), the majority of computational linguistic work on automatic detection of harmful speech thus far fails to capture more subtle forms of harm. My work seeks to address that gap by examining occurrences of tone policing in online spaces, and investigating the feasibility of computational detection of tone policing.

3 Constructing the Dataset

In order to check for the presence of tone policing in textual interpersonal communication, we draw

Figure 1: Tone policing sentences from sociological literature submitted to Mathew et al. (2022) model. Both sentences classified as primarily "normal"

Figure 2: Tone policing sentences from sociological literature submitted to Vidgen et al. (2021) model. Both sentences classified as 100% "not hate"

data from tweets on Twitter.

3.1 Data Collection

Data for this project was collected via the Twitter API v1.1 Academic Research Access tier. Tweets were collected over a period of 11 months in 2021 and 2022, through the Twitter sample streaming endpoint which delivered a random 10% of publicly available English tweets in real time. Due to the deprecation of the Twitter API v1.1 and the removal of the Academic Research tier starting in late 2022, data collection ceased in the winter of 2022.

Several preliminary filters were applied to the dataset. To extract instances of interpersonal communication, tweets were selected in pairs of a reply tweet and its immediate parent, so that the reply tweet would by nature be responding to the parent tweet. Any reply tweets that did not contain the second-person pronoun "you" in any form were discarded, as a broad filter for tweets that did not reference the prior interlocutor in some way, as referencing the prior tweet's author is a prerequisite for tone policing them.[1] Finally, we selected only tweets that were in English and tweets that did not contain non-textual media such as images, videos, or gifs.

4 Rule-Based Lexical Classifier

Due to the very low likelihood of any particular live-sampled tweet containing an instance of tone policing, human annotation of the raw dataset was infeasible at this stage. Because LLMs rely on having annotated training data, we chose to begin with a rule-based classification approach to the raw data, following Sheng et al. (2021)'s identification of salient n-grams for ad hominems in text. In addition to facilitating human annotation, the rule-based approach also confers the advantages of increased interpretability and transparency, compared to typical "black box" style large language models. Therefore the rule-based classifier was constructed with an eye towards the insights into linguistic structures that it can provide.

4.1 Pattern Matching

Because the act of tone policing presupposes a reference to tone at all in the sentence, we drew from naturalistic examples in the sociological literature on tone policing to develop sets of adjectives, adverbs, and verbs that could refer to tone or manner of expression. For example, from the tone-policing sentence "You need to speak in a more respectful manner," we extracted the verb "speak" and the adjective "respectful." We then used GloVe word embeddings to expand these lists by adding words with closely related meanings.

Using the Python library spaCy's DependencyMatcher API, we constructed a series of syntactic dependency patterns based on observed sentence constructions in the sociological literature. After preprocessing, including making all tweets lowercase, removing any text in quotations, and remov-

Category	Example Sentence(s)
Critical Statements	"you (PRON/NP) are (V) so whiny (ADJ/ADV)" "your comment (PRON/NP) was (V) unnecessarily aggressive (ADJ/ADV)"
Suggestions for Change	"you (PRON) should (MODAL) be more civil (ADJ/ADV)"
Rhetorical Questions	"you (PRON) mad (ADJ/ADV) huh? (QUESTION)" "why (QUESTION) are you (PRON) so angry (ADJ/ADV)"
Directives	"(NO PRON) try (V) dressing (V) more appropriately (ADJ/ADV)"
Self-Centering	"I (subj PRON) can't take (V) your voice (obj PRON/NP) seriously (ADJ/ADV)"

Table 1: Typology of linguistic structures observed to characterize sentences engaging in tone policing, with annotated example sentences for each category

Total: 1074 tweets	Tone Policing	Not Tone Policing	Ambiguous
Without Context	271	336	440
With Context	314	656	77

Table 2: Human annotation of data from rule-based classifier, with and without parent tweet context provided.

ing usernames, the classifier was run on the dataset of reply tweets. Several rounds of manual tuning were conducted via a process of human annotation of the classifier results and revision of the dependency patterns based on the annotation. Via this process, we constructed a typology of linguistic structures most often observed to characterize tone policing, laid out in Table 1.

[1] Although it is possible to tone police a third party, using a pronoun other than "you," in this work we chose to follow the Sheng et al. (2021) heuristic and limit our analysis to second-person tone policing because its form is clearer in conversation.

4.2 Annotation Process

The majority of the human annotation at this stage was completed by the author of this paper. Tweets were annotated with one of three labels: -1 ("not tone policing"), 1 ("is tone policing"), and 0 ("ambiguous," i.e. it is not possible to judge whether or not this is tone policing given the available information). Working from our definition of tone policing developed in Section 2.1, we subsequently developed an annotation guideline for the presence of tone policing in tweets. The guideline is as follows:

A tweet is engaging in tone policing if:

- it is a response to a tweet by a different person
- it is implicitly or explicitly critical of the parent tweet's tone or manner of expression
- it implicitly or explicitly dismisses the content of the parent tweet
  - a tweet containing solely an insult to the prior tweet author would not be tone policing unless it is used to imply an invalidation of their message
- the dismissal occurs specifically due to tone or manner of expression of the parent tweet
  - a tweet insulting a characteristic of the parent tweet author would not be tone policing unless it indicates that the characteristic affected or is present in the parent tweet in question

Table 3 contains examples of replies to a parent tweet and their classification as tone policing or not tone policing, as per the annotation guideline above.

As noted earlier, tone policing may also co-occur with social dynamics such as imbalances of power or privilege. However, due to the difficulty of definitively determining identity information on a social media platform such as Twitter, for the current work we choose to focus on the rhetorical aspects of tone policing.

The first stage of the annotation process for manual tuning of the rule-based classifier was completed by referencing only the reply tweets, as the classifier took only that information. Subsequently a second stage of annotation was conducted by referencing both the reply tweet and the parent tweet

Parent Tweet: "I'm so f**ing pissed off that the actor that was my literal childhood crush is apparently homophobic, like what is f**ing wrong with people!"		
Reply	Tone Policing?	Why?
"Uggghh I feel you I hate that"	No	Not critical of parent tweet tone or manner of expression
"You're a moron"	No	Just an insult - not referencing content of message
"Shouldn't an actor be smart enough to not say that shit in front of cameras?"	No	Critical of third party manner of expression, not parent tweet
"You are being way too sensitive he literally just made one joke"	Yes	Dismisses content of parent tweet (anger at homophobic actor) due to criticism of tone ("too sensitive")

Table 3: Examples of reply tweets and their classification as tone policing or not tone policing per the annotation guidelines, and justification thereof

Category	Critical Statements	Suggestions for Change	Rhetorical Questions	Directives	Self-Centering
Proportion	0.856	0.136	0.176	0.001	0.001

Table 4: Proportions of data from rule-based classifier that match each category in our typology of tone-policing tweets.

in order to make a judgment. This dataset was then suitable for the use of an LLM based classifier.

Data from the rule-based classifier after human annotation are summarized in Table 2. Additionally, Table 4 lays out the proportions of data from the rule-based classifier which match each category in our typology of tone-policing tweets. (Tweets may match more than one category).

5 Rule Based Classifier Results

We examine the performance of our rule-based classifier at identifying tone policing.

5.1 Quantitative Results

From a raw dataset of just under 2 million reply tweets, the rule-based classifier identified about 2500 tweets as tone policing. Of these, as summarized in Table 2, 1047 tweets were human-annotated, yielding 271 tweets labeled tone policing, 336 labeled not tone policing, and 440 labeled ambiguous. The classifier therefore achieved a 25.8% precision in terms of matching human judgment on examples that are definitively tone policing. If ambiguous examples are included, then the classifier achieves a 67.9% precision at identifying reply tweets that are either possibly or certainly tone policing.

5.2 Quantitative Results Analysis

The rule-based classifier was run only on reply tweets, and the human annotation accordingly was also done by only considering the reply tweet. The large number of examples human-annotated as ambiguous indicates that both human judgment and the rule-based approach struggle with ambiguous examples when provided only the text of the reply tweet. Per Table 4, a majority of examples matched our typological category of critical statements. This preponderance of critical statements is largely due to their simplicity, i.e. the least restrictive syntax required to match such a construction. Because tweets could match more than one category, many tweets that matched a more complex syntactic category also had a sub-clause that fit into the critical statement category. The relative ubiquity of this type of sentence construction, however, likely contributed to the amount of false positives from the classifier as well as the amount of examples human-labeled as ambiguous. Phrases such as "You're just jealous" or "Why are you still so scared?" are often used in a tone policing context, but are also quite often used humorously or sarcastically, in ways that are not clear from just the text of the reply tweet. Additionally, adjectives and

adverbs like "stupid," "crazy," "foolish," and "mad" (in some dialects of English) are often used in a casual, non-literal manner, although they may also be used to tone police. These usages were difficult to disambiguate without conversational context. However, the usage of adjectives such as "emotional," "harsh," "responsible," and "angry" was much more likely to correlate with the example being classified as unambiguously tone policing.

Given the large number of tweets human-annotated as ambiguous when considering only the reply tweet, the examples were subsequently re-annotated using both the reply tweet and its immediate parent, in order to provide some conversational context. With the inclusion of the parent tweet, among 1047 examples, 314 tweets were human-labeled as tone policing, 656 tweets as not tone policing, and 77 tweets as ambiguous. The addition of context significantly reduced the number of tweets classified as ambiguous, although most of those tweets were then classified as not tone policing.

5.3 Qualitative Analysis

See Table 7 for example tweets referenced here. Many tweets were human-identifiable as tone policing even without parent tweet context, e.g. tweet 245, which contains the sarcastic phrase "Why are you so angry? Did your mom unplug your video games?" Context does help resolve ambiguity in tweets 270 and 243, however. Both of these reply tweets are relatively short and use adjectives like "dumb" or "stupid," which have low specificity and can be used in a myriad of ways, whether jokingly, as a general insult to a person, or as tone policing. Parent tweet context therefore appears to clarify the usage for the human annotator, although in certain cases such as tweet 264, the human annotation changes from tone policing to ambiguous - in this case, because a tweet initially interpreted as tone policing was preceded by an apparently joking comment, making the tenor of the reply unclear. Many examples that remained annotated as ambiguous even with context were shorter both in parent and reply tweets, simply not providing enough information to make a clear decision.

6 LLM Classifier

Because of the nuanced nature of this task, we also investigate a computational model-based approach to identifying tone policing, as transformer attention mechanisms have in recent years demonstrated significant success at capturing contextual information that influences linguistic judgments, even in relatively subtle tasks such as the identification of empathy. (Sharma et al., 2020)

6.1 DistilBERT Model

A commonly used transformer-based model in the field of NLP is BERT (Bidirectional Encoder Representations from Transformers; (Devlin et al., 2019)), often used for tasks such as text classification. Because BERT is optimized for very large datasets, we instead turn to DistilBERT, a smaller and faster version of BERT that is more ideal for smaller datasets due to its stripped-down nature (Sanh et al., 2020)

6.2 Finetuning Process

The DistilBERT model has been pre-trained on large language datasets. In order to apply it to the task of identifying tone policing, we finetune the model on our own tone policing data, adapting code from Valencia (2023) used to finetune DistilBERT on hate speech data.

Our annotated dataset is split into training, validation, and test sets with a ratio of 70:15:15. The model is then trained for 5 epochs using a learning rate of 5e-5 and a batch size of 16. Hyperparameters were selected via a grid search as well as via other literature on finetuning BERT-type models for smaller datasets.

Depending on the task, the model was run on either the full dataset or a subset. Table 5 lays out the data on which each task was performed and the label distribution for each.

6.3 Task: Three Class Prediction

Since our data is annotated with three labels (tone policing, not tone policing, ambiguous), we first examined whether the DistilBERT classifier could accurately predict all three labels. This task is most akin to a real-world setting, as naturalistic data is not necessarily easily interpretable by humans. The model is provided with the text of the parent and reply tweets in the form "PARENT [SEP] REPLY", i.e., the parent tweet concatenated with the reply tweet with a separator token between the two. We also evaluated the classifier's performance when not given the parent tweet as context, in which case the text provided to the model was simply the reply tweet.

Dataset	Label Distribution (% tone policing/non tone policing/ambiguous)	Task(s)
Full	0.30/0.63/0.07	3 class prediction
No ambiguous	0.32/0.68/0	2 class prediction
Precision oriented	0.30/0.70/0	Precision oriented 2 class prediction
Recall oriented	0.37/0.63/0	Recall oriented 2 class prediction

Table 5: Datasets used for each task and associated label distribution

6.4 Task: Two Class Prediction

In the three class prediction framework, predicting which tweets are ambiguously tone policing is likely a more complex task than predicting which tweets definitively are or are not tone policing. For this reason, our second task evaluates the classifier's performance with a two-class framework, i.e., predicting only whether a tweet is or is not tone policing. Tweets annotated as ambiguous are removed from the data provided to the model. This task provides a best-case-scenario for model performance; assuming a completely unambiguous world, we can therefore establish an upper bound for prediction accuracy. Performance is again evaluated both with and without the parent tweet text included as context.

6.4.1 Subtasks: Precision and Recall Oriented Two Class Prediction

With an eye to downstream applications, we also assess the performance of precision-oriented and recall-oriented model versions. For these versions, examples labeled as ambiguous were included in the dataset, but for the precision-oriented version their labels were changed to -1 ("not tone policing"); for the recall-oriented version, to maximize the classifier's sensitivity, ambiguous examples were changed to 1 ("tone policing").

7 LLM Classifier Results

We examine the effectiveness of our LLM classifier at identifying tone policing. Statistical significance for the LLM results is calculated via a bootstrap sampling significance test (Fornaciari et al., 2022)

7.1 Quantitative Results

The DistilBERT model was trained using the re-annotated dataset and then applied to the tasks of three-class prediction (including ambiguous examples) and two-class prediction (excluding ambiguous examples). In order to determine the importance of context in the model's performance, for each task, the model is run both with and without the parent tweet text included as context. The most common class (MCC) baseline is computed for each task. The results are summarized in Table 6. The model achieves a significant improvement over baseline in both the three class prediction task and the two class prediction task. In the two class prediction task, the addition of context does not improve the model's accuracy on the test set, and the hypothesis that the context-added version improves on the without-context version does not reach significance ($p < 0.372$). The three class prediction task shows a slight numerical improvement in accuracy with context, but this improvement does not reach significance ($p < 0.441$).

In the precision oriented two class subtask, for which ambiguous examples were labeled -1, the model achieved an accuracy of 92.36%. In the recall oriented version, for which ambiguous examples were labeled 1, the accuracy was 87.26%.

7.2 Quantitative Results Analysis

The model achieves large improvements over baseline in both tasks, indicating that it is relatively successfully able to classify examples as tone policing or not tone policing. As predicted, accuracy is higher on the two class task, which removes the ambiguous examples to assess performance in a less naturalistic "ideal world" scenario. The model however does still perform relatively well on the three class task, indicating some ability to correctly predict ambiguous examples. In both tasks, the inclusion of some conversational context via the text of the parent tweet does not lead to an improvement in performance. In contrast, during the

Version	Three Class Prediction Accuracy	Two Class Prediction Accuracy	Two Class Precision Oriented	Two Class Recall Oriented
With Context	85.99	87.59	92.36	87.26
Without Context	85.35	87.59	N/A	N/A
Baseline	58.60	63.45	70.19	62.72

Table 6: % accuracy of DistilBERT model performance on two class (excluding ambiguous) and three class (including ambiguous) classification of tone policing, as well as the most common class baseline for each task

second stage of the human annotation process when provided parent tweet context, quite a few tweets previously labeled ambiguous were instead classified as tone policing or not tone policing. This indicates that context has a significant impact on human ability to give concrete judgments of tone policing, but does not play a role in the LLM's overall accuracy.

Figures 3 and 4 show the confusion matrices for the model's performance on each task with and without context, normalized against true labels. For the two class task, the accuracy of classifying non tone policing examples is quite high, while the accuracy for tone policing examples is lower but still good. The confusion matrices are quite similar between settings, showing that when context is added, there is only a slight increase in tone policing examples classified correctly, but a slight decrease in non tone policing examples classified correctly.

For the three class task, without context, the model is quite successful at predicting examples labeled -1 and 1, but the majority of the ambiguous examples are predicted incorrectly. The addition of context causes a slight improvement in prediction of non tone policing examples, but a decrease in accuracy of classifying tone policing examples: more of these are predicted to have the label -1 instead. However, the performance on ambiguous examples is significantly improved. Almost no ambiguous examples are misclassified as tone policing, while without context, 44% of ambiguous examples were misclassified as tone policing, indicating that when context is added, the model may be shifting weight from the tone policing label to the ambiguous label.

Figures 5 and 6 show for the examples predicted to be each label, the mean and standard deviation of the probability calculated, i.e., how certain the model is about assigning the label in question. In the two class task, the mean probability for the non tone policing examples is higher than for the tone policing examples, indicating that the model is more confident in predicting examples labeled -1. The mean probabilities are very similar between the context and no context settings, but with context, the standard deviation for examples labeled 1 has increased slightly, indicating that the addition of context has actually made the model slightly less confident about tone policing examples. For the three class task, providing context increased the mean probability for all three labels, most dramatically in the case of ambiguous examples. Added context therefore seems to allow the model to be both more certain and more accurate about predicting ambiguous examples. However for the label 1, the probability is quite high across both settings despite the decrease in accuracy with context, indicating possible false confidence in those predictions with context.

7.3 Qualitative Analysis

See Table 7 for example tweets referenced here. All versions of our models are quite good at predicting non tone policing examples both with and without context, e.g. tweets 269 and 270, despite 270 being a case in which the human annotator required context. Our two class model's performance is improved slightly (from 70 to 75%) when provided context. Context appears to have helped reach the correct judgment when both parent tweet and reply tweet have similarities in topic: in tweet 42, the mention of "religion" in the parent tweet can be linked to "scared" and "hell" in the reply tweet, while in tweet 245, the parent tweet uses offensive language ("shit," "morons") which provides a clear target for the reply tweet tone policing of "anger." On the other hand, the parent tweet for tweet 164 provides very little information, so the model's judgment is unchanged between the two settings.

The three class model performs relatively well (89%) on tone policing examples without context. This is improved performance over the two class

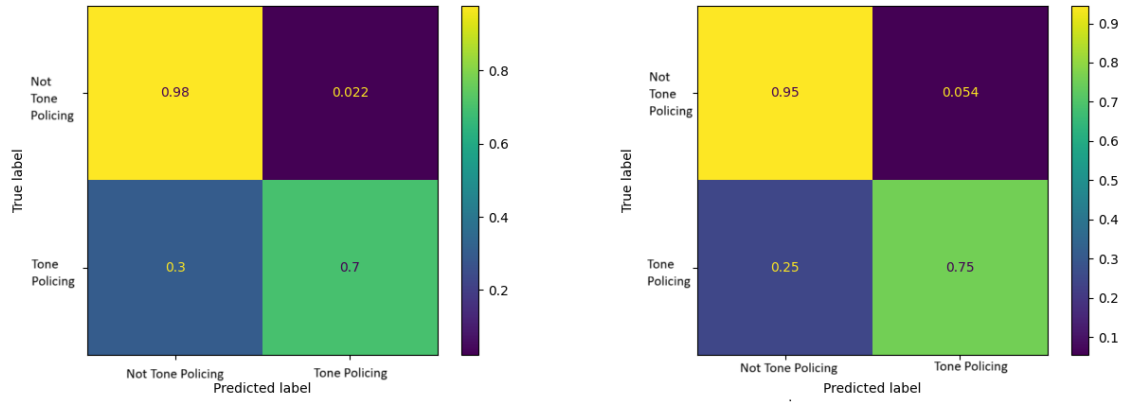

Figure 3: Confusion matrices for model run on two-class task, without context (left) and with context (right). This shows the proportion of examples that were predicted to have each label, normalized to true labels.

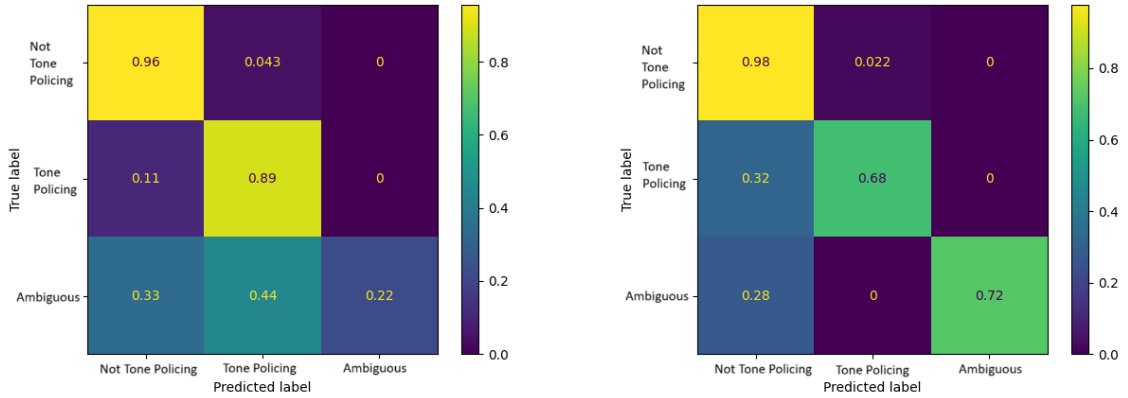

Figure 4: Confusion matrices for model run on three-class task, without context (left) and with context (right). This shows the proportion of examples that were predicted to have each label, normalized to true labels.

model, which may simply be due to the model having more examples on which to train with the ambiguous examples kept in the dataset. However, including context decreases accuracy on tone policing tweets while significantly increasing accuracy on ambiguous tweets. A common correlate with human annotation of ambiguity is both parent tweet and reply tweet being quite short, and the three class model with context does well on these, e.g. tweets 263, 264, 265, and 43 which are all correctly identified as ambiguous. However it struggles with tweet 166 which has longer text in the parent and reply tweets. With examples that are tone policing, we hypothesize that the model with context may struggle due to the stylistic differences from the different tweet authors; it performs best on tweets such as 271 and 164, where the parent tweet is relatively short and the reply tweet is relatively long.

8 Further Discussion

Having analyzed the performance and errors of our models, we turn to the broader scope of insights our models have provided about tone policing, and return to the research questions posed in Section 1.

8.1 RQ1 - What linguistic structures characterize tone policing?

Via the sociological literature and the results of our rule-based classifier, we developed a typology of linguistic structures that characterize tone policing, i.e., critical statements, suggestions for change, rhetorical questions, directives, and self-centering (Table 1). Of these, critical statements are most common in our dataset, followed by rhetorical questions. We find that tone policing is also commonly characterized by references to emotion or emotional expression, e.g. “why are you angry?” “you’re too sensitive” “because you are scared”

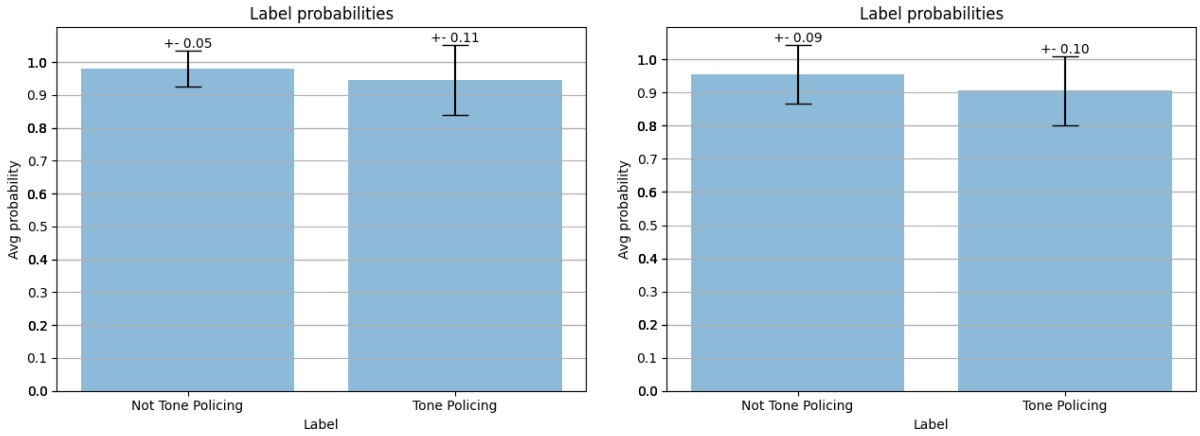

Figure 5: Bar charts for model run on two-class task, without context (left) and with context (right). This shows the mean and standard deviation of the model-calculated probability for examples predicted to have each label.

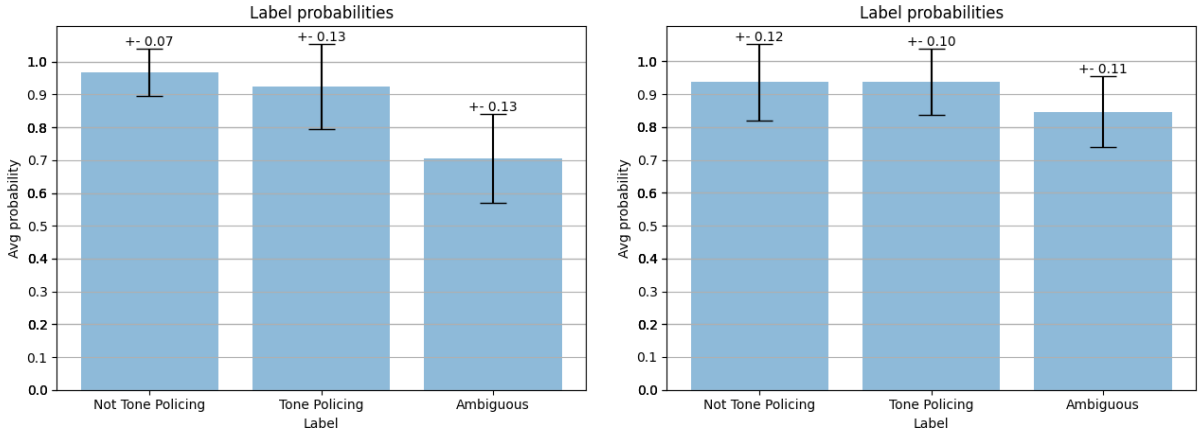

Figure 6: Bar charts for model run on three-class task, without context (left) and with context (right). This shows the mean and standard deviation of the model-calculated probability for examples predicted to have each label.

"you're just jealous" from Table 7. This is corroborated by studies such as Nuru and Arendt (2019) and Biddle and Hufnagel (2019), which emphasize that "too-strong" emotional expression from a speaker leads to discomfort, which then prompts the listener to tone police the speaker. From the results of our computational models, we can also conclude that tone policing rarely occurs as short, standalone utterances; instead, it is often paired with some amount of justification or elaboration, e.g. "you're just jealous because [...];" "stop whining about your problems and instead [...]". This justification or elaboration could potentially in fact be more predictive of tone policing than the presence of parent tweet context. We hypothesize that tone policing via more detailed, substantive replies may better achieve the definitional aim of dismissing or invalidating the original tweet's content: shorter replies cross into the realm of basic insults, e.g. "you're so dumb" with no further context, and therefore may be more readily ignored entirely by the parent tweet author.

8.2 RQ2 - Can a computational model accurately detect tone policing in text-based conversations?

Of the versions of our models tested, the three-class model without parent tweet context performed best at detecting instances of tone policing (89% accuracy), and all models overall performed better than baseline. This result shows that despite tone policing's status as a relatively subtle type of potentially harmful speech which humans can sometimes have difficulty classifying without context, computational models trained on human judgments can match human performance even without having been provided parent tweet context. While our three-class model with parent tweet context per-

Tweet #	Parent Tweet	Reply Tweet	2NC	2C	3NC	3C	HNC	HC
269	If you are stupid in Africa you will be famous. But if you are positive your struggle continues. Why so #Africa	@USERNAME Did you say stupid...how many stupid people are famous, mention them	0	0	0	0	0	0
270	@USERNAME @USERNAME But her mom is shareholder so	@USERNAME @USERNAME @USERNAME You don't even know her mom name lol, dumb just as always	0	0	0	0	2	0
243	@USERNAME Your audience and 'the internet audience' are the same people Vince, it's goddamn 2022.	@USERNAME @USERNAME Do you realize how stupid of a statement that is?	0	1	0	2	2	0
271	@USERNAME @USERNAME @USERNAME Take your meds old man.	@USERNAME @USERNAME @USERNAME Well, now you're just being rude. And that suggests that you don't have a sound argument anymore. In fact, as I look back over this discussion with you, I'm not sure I see one of those. I'll let you have the last word if you want it, but I think maybe I haven't lost after all.	1	1	1	1	1	1
46	@USERNAME Inaaa, still counts for me	@USERNAME You're too sensitive to spoilers...everyone is but this is minor	1	1	1	0	1	1
245	@USERNAME @USERNAME @USERNAME @USERNAME @USERNAME Nope. Just sick of explaining shit to morons.	@USERNAME @USERNAME @USERNAME @USERNAME @USERNAME See, now that is showing anger. Don't be angry. Why are you so angry? Did your mom unplug your video games?	0	1	1	0	1	1
42	@USERNAME You have lost the sense. I am not a perfect Muslim but I will never question my religion at all...	@USERNAME @USERNAME because you are scared of imaginary hell fire	0	1	1	0	1	1
164	@USERNAME @USERNAME @USERNAME @USERNAME Okay pedophile	@USERNAME @USERNAME @USERNAME @USERNAME You're just jealous because Biden defeated tRump's virus, and is showing worldwide leadership through his flawless handling of Russia. And he kept tRump's promise to get out of Afghanistan. He's also masterfully tackling this stifling inflation. Cleaning up tRump's messes.	0	0	1	1	1	1
263	@USERNAME Lmao stop trying to be the victim.	@USERNAME @USERNAME How you broke and dumb? Pick a struggle.	0	0	1	2	2	2
264	@USERNAME it's true- i'm jealous	@USERNAME you shouldnt be jealous	0	0	0	2	0	2
265	@USERNAME Gtfoh	@USERNAME You're always angry	0	0	2	2	2	2
43	@USERNAME Congratulations	@USERNAME @USERNAME Are you being foolish right now or something? Is this even remotely funny?	1	0	1	2	1	2
166	@USERNAME So you don't think that 200lbs of fatt your packing around is affecting your mental state? Maybe you should go for a walk	@USERNAME @USERNAME As long you're helping people find THEIR (fixed it for you, so you don't look dumb and unable to spell on your bio) best self. If you prop yourself up as a person of positivity then jab at weight, maybe you aren't doing Jack shit more than being rude?	0	0	1	0	2	2

Table 7: Sample tweets from dataset with the label predicted by each version of DistilBERT model as well as human (true) label. Key: 2NC = 2 class model with no context; 2C = 2 class model with context; 3NC = 3 class model with no context; 3C = 3 class model with context; HNC = human label with no context; HC = human label with context (True Label)

forms relatively well at predicting tweets which human annotators labeled ambiguous, it may be that this is not a useful category of example to predict, as it seems to come at a distinct tradeoff with accuracy at detecting tone policing tweets. It is also possible that providing context via the full thread of tweets preceding the tweet in question may improve performance; however, this is a more complex task, as it can potentially include tweets by many different authors.

Our results also show promise for downstream applications; our recall-oriented model, which maximizes detection sensitivity, can guide deeper investigation of tone policing, including any edge cases in terms of linguistic structure or word usage not caught by our current models. Our precision-oriented model, on the other hand, can serve as a foundation for automated moderation systems in the real world. Maximizing precision means that the system will only flag and intervene in examples that it is most certain are tone policing, a crucial aspect of auto-moderation systems that are seen as trustworthy by users.

8.3 RQ3 - Are certain classes of linguistic structures more or less likely to be detectable as tone policing by a computational model?

The combination of our rule-based classifier and our LLM classifiers allows us to mitigate some of the interpretability issues inherent to "black box" machine learning models. Examining our results and qualitative analyses through the lens of our typology of tone policing, we see from Table 8 that on a broad level, the LLM models appear to achieve good performance with all sentence constructions laid out in our typology. As discussed earlier, critical statements, i.e. direct comments on tone, make up the largest proportion of our data. Thus, the performance of our models on this category of data is similar to performance on the overall dataset, with the two-class model achieving higher accuracy than the three-class model. Rhetorical questions comprise the second largest category, and in this case, the two class model with context achieves the highest accuracy. Tone policing through rhetorical questions often occurs in the form of snarky or sarcastic comments - e.g. "you angry?" "why are you so sensitive?" However in some cases, genuine questions ("why would you say that?") can be difficult to disambiguate from snarky rhetorical

questions ("why would you say that?") This category of sentence inherently has some ambiguity, but it is likely that the addition of the ambiguous label in the three-class model reduced the model's ability to correctly positively classify tone-policing sentences rather than improving accuracy overall, as can be seen in the confusion matrices in Figures 3 and 4. Providing parent tweet context, however, does improve accuracy across both models, likely by contributing to more effective disambiguation of genuine vs. rhetorical questions.

Of the remaining three categories in our typology, their relative scarcity in our dataset unfortunately prevents a substantive quantitative analysis of model performance. Broadly, performance on suggestions for change, directives, and self-centering appears invariant across models. Qualitatively, statements made as suggestions for change typically occur either as part of arguments filled with other inflammatory language ("you really ought to keep quiet"), or as instructive comments ("you should try speaking more clearly to your Alexa"). Thus, it is reasonable to deduce that the models do not have significant difficulty differentiating inflammatory suggestions from instructive ones, regardless of parent tweet context. The directive sentence construction is simply very uncommon in the brief, casual style of conversation typical to Twitter. The self-centering category also yielded very few matches in our data, although this is likely due to the more restrictive dependency patterns necessary for our rule-based classifier to match such a construction while excluding non-tone-related self-centering commentary.

Overall, one source of error is very short reply tweets, especially ones which refer to the parent tweet with words like "dumb," "crazy," "jealous," etc. These tweets typically occur as critical statements, e.g. "you're just jealous;" "you sound so dumb." Because these words lack specificity in casual usage, it may be difficult to distinguish humorous or sarcastic usage from actual tone policing. Our models appear to perform better on tweets with more specific, and therefore more easily interpretable, emotion words, although a full analysis on this hypothesis has not yet been completed. Our models also appear to more accurately detect tone policing as part of longer tweets, in which the user often elaborates on their tone policing and provides their own contextual information. Future work is planned to expand and restructure our rule-based classifier to acquire more data on our least common typological categories, as well as to examine in more detail the relationship between model performance and the usage of particular emotion words.

Category	Critical Statements	Suggestions for Change	Rhetorical Questions	Directives	Self-Centering
3 Class Model With Context					
# Correct	132	15	27	2	6
# Incorrect	22	3	5	0	0
Accuracy	0.857	0.833	0.844	1.0	1.0
3 Class Model No Context					
# Correct	131	15	26	2	6
# Incorrect	23	3	6	0	0
Accuracy	0.851	0.833	0.813	1.0	1.0
2 Class Model With Context					
# Correct	119	9	13	0	0
# Incorrect	18	0	2	0	0
Accuracy	0.869	1.0	0.866	N/A	N/A
2 Class Model No Context					
# Correct	120	9	12	0	0
# Incorrect	17	0	3	0	0
Accuracy	0.876	1.0	0.80	N/A	N/A

Table 8: # of tweets classified correctly and accuracy per category in typology of tone policing tweets, for each LLM model variant

8.4 Conclusion

Via answering these research questions, our work contributes a clearer understanding of the linguistic characteristics of tone policing, as well as evidence that large language models can indeed be trained to correctly identify subtler forms of harmful speech. Our results indicate that our model can also be a robust foundation for future moderation tools for online communities, as well as for deeper investigation of the dynamics of power, oppression, and other social factors surrounding tone policing.

9 Limitations and Future Work

In addition to what has previously been discussed, some limitations of our current work derive from our choice of data. Because we used conversations on Twitter, interlocutors typically produce short, relatively stylistically casual utterances. We also excluded tweets with non-textual modalities, such as picture or video attachments, and limited the scope to English-language tweets for ease of annotation. However, as socially fraught as tone policing is, it may manifest in fairly different ways if studied in longer-form conversations, in different modalities, or in other languages; our work should

not be taken to generalize across all contexts. Our work also considered only direct tone policing of one's interlocutor, rather than of a third party; future work that studies third-person tone policing may find that it has very different linguistic characteristics.

In our future work on this topic, we plan to address the question of cross-cultural applicability by soliciting crowdsourced annotations of our data. Ideally, by hiring a pool of socially and culturally diverse annotators, our ground truth labels of our examples will be made as robust as possible. When crowdsourcing judgments of tone policing, we plan to also seek more granularity of annotation by providing our annotation guideline and having annotators rate examples on a Likert scale from "definitively tone policing" to "definitively not tone policing," allowing them to record the degree of ambiguity which they think pertains to any example. We predict that this may allow our model to learn to differentiate ambiguous examples from tone policing with more confidence.

Finally, this work was largely centered around the problem of computationally detecting tone policing from text. However, tweets on Twitter also typically contain some amount of social or demographic information, via user-populated fields such as username, bio/description, location, and so on. Given that sociological work on tone policing emphasizes the roles that privilege and power play, future work that integrates social information into analyses of tone policing will likely be significantly helpful in advancing sociolinguistic understandings thereof. Sociolinguistically oriented research questions that could be investigated by such work include: whether certain users engage in tone policing more often than others or are tone policed more often than others, depending on their social characteristics; which and how many emotion words are used and by which people; or whether certain topics of conversation are more likely to lead to instances of tone policing and why.

10 Ethics Statement

Because our dataset was drawn from public tweets on Twitter, we did not have to gain permission from individual users to include their tweets. Nevertheless, we understand that many users do not read social media user agreements that include clauses about their data being used for research purposes. For this reason, we did our best to ensure privacy of the users, including masking usernames in tweets and not providing user information during the annotation process or in this article. We are open to any feedback on how to further improve ethical considerations in this aspect of our work.

In future iterations of this work, as discussed in our Limitations and Future Work section, we plan to hire crowdsourced annotators from diverse backgrounds in order to assure more culturally balanced judgments of tone policing. Given that some tweets in our dataset contain content that is toxic or sexual in nature, we also plan to develop guidance regarding how best to avoid such messages causing distress to annotators.

The broader impact of our work centers on more accurately being able to identify a subtle, yet still harmful, rhetorical strategy. Sociological work indicates that tone policing is often aimed at minorities who are already disproportionately silenced, and that it leverages power dynamics in order to perpetuate societal standards of privileged groups and force others to conform to them. Therefore we hope that our work's contribution towards linguistic understanding and computational identification of tone policing is a step towards mitigating the harm that it can cause and the systems of power and privilege that enable it.

References

Catharine Biddle and Elizabeth Hufnagel. 2019. Navigating the "Danger Zone": Tone Policing and the Bounding of Civility in the Practice of Student Voice. *American Journal of Education*, 125(4):487–520. Publisher: University of Chicago Press ERIC Number: EJ1223294.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. ArXiv:1703.04009 [cs].

Patricia G. Davis. 2020. Spatiality at the Cyber-Margins: Black-Oriented Blogs and the Production of Territoriality Online. *Social Media + Society*, 6(2):2056305120928506. Publisher: SAGE Publications Ltd.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [cs].

Miranda Flores. 2016. Stealing the Mic: Struggles of the Black Female Voice in Rap. *Undergraduate Honors Thesis Collection*.

Tommaso Fornaciari, Alexandra Uma, Massimo Poesio, and Dirk Hovy. 2022. Hard and Soft Evaluation of NLP models with BOOtSTrap SAmpling - BooStSa. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 127–134, Dublin, Ireland. Association for Computational Linguistics.

Marc Lamont Hill. 2018. "Thank You, Black Twitter": State Violence, Digital Counterpublics, and Pedagogies of Resistance. *Urban Education*, 53(2):286–302. Publisher: SAGE Publications Inc.

Maisha Z. Johnson. 2015. What We Can All Learn From Nicki Minaj Schooling Miley Cyrus on Tone Policing.

Joseph Kwarteng, Serena Coppolino Perfumi, Tracie Farrell, and Miriam Fernandez. 2021. Misogynoir: public online response towards self-reported misogynoir. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 228–235, Virtual Event Netherlands. ACM.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2022. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. ArXiv:2012.10289 [cs].

Audra K. Nuru and Colleen E. Arendt. 2019. Not So Safe a Space: Women Activists of Color's Responses to Racial Microaggressions by White Women Allies. *Southern Communication Journal*, 84(2):85–98.

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label Categorization of Accounts of Sexism using a Neural Framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1642–1652, Hong Kong, China. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv:1910.01108 [cs].

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. ArXiv:2009.08441 [cs].

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. "Nice Try, Kiddo": Investigating Ad Hominems in Dialogue Responses. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 750–767, Online. Association for Computational Linguistics.

Luis Valencia. 2023. Fine-tuning DistilBERT with Your Own Dataset for Multi-classification Tasks.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. ArXiv:2012.15761 [cs].