Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Special Track on AI and Social Good

Towards a Bipartisan Understanding of Peace and Vicarious Interactions

Arka Dutta[1] , **Syed Mohammad Sualeh Ali**[1] , **Usman Naseem**[2] and **Ashiqur R. KhudaBukhsh**[1]
[1]Rochester Institute of Technology
[2]Macquarie University
ad2688@rit.edu, sa4915@rit.edu, usman.naseem@mq.edu.au, axkvse@rit.edu

Abstract

Human input plays a critical role in modern AI systems. As machines take on increasingly nuanced tasks, it becomes essential for the community to embrace subjectivity and diverse perspectives. However, research on sensitive topics often fails to incorporate diverse and balanced perspectives. This paper makes a key contribution to participatory AI design in the context of conflicts between nuclear adversaries (India and Pakistan); where disagreement between stakeholders is anticipated. The paper explores the notion of *hope speech* detection – detecting de-escalating content in the context of nuclear adversaries on the brink of war – through the lens of participatory AI design and vicarious interactions. We release a dataset of 10,081 social web posts annotated by raters from India and Pakistan and examine the bipartisan nature of the language of de-escalation. Our study reveals that vicarious perspectives can be useful for modeling out-group preferences.

1 Introduction

From traditional supervised machine learning solutions [Mitchell, 1997] to recent RLHF frameworks [Ouyang *et al.*, 2022], human input plays a critical role in most AI systems. However, human disagreement in annotation has long been treated as a by-product of poor annotation study design or human errors while annotating that need to be "corrected" [Pavlick and Kwiatkowski, 2019]. As growing literature suggests that many tasks may not have a single ground truth [Plank, 2022] and label variations may stem from multiple plausible answers or innate subjectivity [Passonneau *et al.*, 2012; Pavlick and Kwiatkowski, 2019; Nie *et al.*, 2020; Jiang and Marneffe, 2022; Deng *et al.*, 2023], human-centered AI faces a moment of reckoning with human subjectivity.

While recent studies acknowledge that rater subjectivity can stem from political leanings [Sap *et al.*, 2022; Weerasooriya *et al.*, 2023] or other demographic factors [Pei and Jurgens, 2023], very few annotation studies incorporate participatory designs [Harrington *et al.*, 2019] and seek balanced representation among annotators. For instance, barring a few studies in the literature (e.g., [Johnson and Goldwasser, 2018; Weerasooriya *et al.*, 2023]), most datasets on US political discourse lack balanced participation from liberals and conservatives. Such balance could be critical for certain sensitive and important tasks such as conflict resolution and peace negotiations. This paper[1] examines a highly sensitive task where ignoring any stakeholders may result in a seriously limited AI system: *the detection of de-escalating social web posts in the context of nuclear adversaries on the brink of war.*

Palakodety *et al.* [2020a] posited that when nuclear adversaries are on the brink of war, detecting and surfacing peace-seeking, hostility-diffusing social media content may have societal benefits. To this end, as a part of the broader literature on counter speech [Benesch *et al.*, 2016], Palakodety *et al.* [2020a] introduced *hope speech* as de-escalating social media content during near-conflict scenarios. The authors curated a *hope speech* dataset of 10,081 YouTube comments annotated by expert social scientists in the context of the 2019 Pulwama conflict between India and Pakistan and demonstrated reliable in-the-wild performance. With a view toward human-centered AI, including diverse and balanced perspectives, we seek to investigate the following.

RQ1: *How aligned are Indian and Pakistani raters on what is de-escalating?* Is it possible that annotators from these two countries have irreconcilable differences in what they evaluate as de-escalating? India and Pakistan are nuclear adversaries and have a shared history of decades of unrest and four major wars [Malik, 2002; Schofield, 2010; Bose, 2009; Staniland, 2013]. A recent study reported a grim forecast of 100 million deaths should there be a full-fledged war between these two nuclear powers [Toon *et al.*, 2019]. Human-centered AI research will benefit from balanced participation from key stakeholders on sensitive tasks. Our paper marks an attempt towards that. We conduct a large-scale annotation study both in Pakistan and India to examine bipartisanship in the language of de-escalation. To our knowledge, no such study exists let alone on a sensitive and important task as detecting *hope speech*.

RQ2: *How well does a Pakistani (Indian) rater predict if a social media post will be deemed as de-escalating by an Indian (a Pakistani) rater?* In our bid to understand biparti-

[1]This paper contains sensitive content.

sanship in the language of de-escalation, we incorporate vicarious interactions in our annotation framework. Recently proposed in Weerasooriya *et al.* [2023] in the context of offensive language in the US political discourse, vicarious interactions measure a rater's ability to represent the values and opinions of those with different perspectives.

Our contributions are the following.

• ***Framework****:* We present a novel lens to examine bipartisanship in annotating de-escalating data leveraging vicarious interactions, integrating human-centered AI. In light of social media's growing significance in understanding and analyzing modern conflicts [Zeitzoff, 2017], and considering that we are currently amid two significant ongoing wars [Samuel, 2023; D'Anieri, 2023], our study contributes to the timely and important topic of bipartisanship in de-escalation.

• ***Social****:* To our knowledge, this study marks the first attempt to investigate bipartisanship in identifying de-escalating content in the context of nuclear adversaries on the brink of war. With a view towards human-centered AI incorporating diverse and balanced perspectives, beyond conducting perhaps the first large-scale annotation study involving more than a thousand Indian and Pakistani raters, our study also involves researchers from both countries.

• ***Resource****:* We present a dataset [2] of 10,081 social web posts (YouTube comments) annotated by 1,639 unique Indian and 1,687 Pakistani raters both for first-person and vicarious perspectives on de-escalation. Each comment is annotated by five Indian and five Pakistani raters. While the web manifestation of modern conflict is an important domain to study, existing supervised solutions have yet to incorporate a balanced and diverse perspective by involving major stakeholders at the table. Our study addresses this gap in the literature.

2 Background

2.1 2019 India-Pakistan Pulwama Conflict

India and Pakistan are nuclear adversaries with four major wars and many skirmishes to date. Kashmir has been a key factor in this continued unrest for decades [Malik, 2002; Schofield, 2010; Bose, 2009; Staniland, 2013]. Overall, an estimated 27,650 soldiers were killed and thousands wounded in these four wars. The 1971 war was the goriest (11,000 killed from both sides) which resulted in the largest number of prisoners of war (90,000 POWs) since the Second World War [Ali, 1983].

The most recent near-conflict situation arose following a terrorist attack in Pulwama, India that claimed the lives of 40 Indian Central Reserve Police Force (CRPF) personnel. This incident followed sharp diplomatic escalations and India reported a major airstrike inside Pakistan's territory (Balakot) and the Pakistani military captured an Indian fighter pilot (Abhinandan Varthaman). When the two countries came precariously close to declaring a full-fledged war, the then Pakistani Prime Minister Imran Khan agreed to release the captured Indian fighter pilot as a gesture of peace.

The social web manifestation of this conflict has been studied by Tyagi *et al.* [2020] (Twitter) and Palakodety *et al.* [2020a] (YouTube). Tyagi *et al.* [2020] looked at the network polarization while Palakodety *et al.* [2020a] analyzed the discourse through the lens of *hope speech* described next.

[2] The dataset is available at https://github.com/ArkaDutta-007/Vicarious-Hope-Speech.

2.2 Hope Speech Detection

To study annotation subjectivity and bipartisanship in high-stakes scenarios, we focus on the prediction task of *hope speech* detection, first proposed in Palakodety *et al.* [2020a] in the context of online discussions relevant to the 2019 India-Pakistan conflict. Aimed at diffusing hostility, a *hope speech* classifier is a nuanced classifier (Palakodety *et al.* [2020a] contain operationalizing definition with illustrative examples) to detect content that contains a unifying message focusing on the war's futility, the importance of peace, and the human and economic costs involved, or expresses criticism of either the author's own nation's entities or policies, or the actions or entities of the two involved countries.

This line of work initiated follow-on research focusing on multilinguality [KhudaBukhsh *et al.*, 2020; Hande *et al.*, 2021] and broader definitions of *hope speech* [PK *et al.*, 2021; Palakodety *et al.*, 2020b; Yoo *et al.*, 2021]. For instance, Chakravarthi and Muralidharan [2021] expanded this term to a broader scope of speech championing equality, diversity, and inclusion and *help speech* detection task is defined in the context of detecting supportive content for disenfranchised minorities [Palakodety *et al.*, 2020b]. The task of *hope speech detection* falls under the broader literature of counter speech [Benesch, 2014; Mathew *et al.*, 2019; Hengle *et al.*, 2024; Saha *et al.*, 2022; Gupta *et al.*, 2023].

2.3 Vicarious Interaction

Traditional hate speech literature asks raters questions about the direct, first-person perception of offense (i.e., *Do you find this social post offensive?*) (see, e.g., [Rosenthal *et al.*, 2021; Mathew *et al.*, 2021]). Weerasooriya *et al.* [2023] recently posed a simple yet powerful and unasked question: *"Do you think group $\mathcal{A}$ will find this social web post offensive?"* The authors introduced this perception of indirect offense as *vicarious offense*. This research resulted in a dataset: `VOICED` consisting of 2,310 social web posts; each post is annotated with at least six Democrats, Republicans, and Independents per post [Weerasooriya *et al.*, 2023].

The study revealed that what $a \in \mathcal{A}$ finds offensive and what $b \in \mathcal{B}$ *thinks* $\mathcal{A}$ would find offensive often do not align. The study further revealed that hot-button issues (e.g., reproductive rights, gun control/rights) influenced raters' ability to predict *vicarious offense*. Also, certain annotators were better at predicting vicarious offense than others. Follow-on research along this line has investigated rater cohesion through the lens of intersectionality [Pandita *et al.*, 2024] and rater consistency [Dutta *et al.*, 2025].

In this paper, we extend the framework of vicarious offense to de-escalation. Beyond asking a rater if they find a social web post de-escalating, we also ask a Pakistani (Indian) rater if they think an Indian (Pakistani) rater would find the content de-escalating. To our knowledge, vicarious interactions have only been studied in the context of online toxicity in US political discourse and never in the context of de-escalation.

Our work examining the bipartisanship of de-escalating language through the lens of vicarious interactions thus breaks new ground.

3 Annotation Study Design

Palakodety *et al.* [2020a] curated a dataset of 10,081 YouTube comments on relevant videos to study de-escalation. Each instance in this dataset is labeled as *hopeSpeech* or *notHopeSpeech*. The annotation guidelines follow the detailed rubric of *hope speech* described in Palakodety *et al.* [2020a] and the annotations were conducted by expert social scientists.

To have a more diverse and balanced public perspective on de-escalation, unlike Palakodety *et al.* [2020a], we consult crowd workers from India and Pakistan. Instead of presenting a complex operationalizing definition of *hope speech*, we simplify the annotation task to reduce cognitive load on them. Keeping rater instructions simple to reduce cognitive load on the raters is a recommended practice in crowdsourcing literature [Finnerty *et al.*, 2013]. We present the raters with a brief description of the context and instruct them to label each instance as *de-escalating* or *Not-de-escalating*. The label *de-escalating* maps to *hopeSpeech*, and the label *Not-de-escalating* maps to *notHopeSpeech*. Following Sap *et al.* [2022] and Weerasooriya *et al.* [2023], we collect standard, coarse-grained demographic information related to age and gender.

To avoid sampling bias, we re-annotate the same dataset with raters from India and Pakistan. All of these instances are authored in English with some degree of code switching which is consistent with datasets from the Indian subcontinent [Sarkar *et al.*, 2020a; KhudaBukhsh *et al.*, 2020]. Our focus on English is grounded in prior literature that indicates that English is one of the primary languages of social web discourse in the Indian subcontinent [Palakodety *et al.*, 2020a; Sarkar *et al.*, 2020b; Chandra *et al.*, 2021].

In what follows, we describe some of the key choices we made for our annotation study.

Crowdsourcing Platform

Very few crowdsourcing platforms host participants from both countries. We found Clickworker to be one such platform that has a reasonable presence in both Pakistan and India. For our survey, we host the data on Qualtrics, making it visible *only* to the residents of India and Pakistan.

Defining De-escalating Content
Expresses positive feelings, unity, and a desire for peace between conflicting parties, fostering understanding and cooperation.
What is NOT De-escalating Content
Lacks positive feelings, impartiality, or promotes harm, contributing to tension and hostility between conflicting parties.

Table 1: Rater instructions about de-escalating and Not-de-escalating content.

Rater Instructions

We first present the rater a brief description of de-escalating content as shown in Table 1. For every instance the rater annotates, these definitions always appear in the preamble.

A sample prompt for a given comment (⟨`comment`⟩) from the Indian survey is as follows:

> **Q1:** *Assume that India and Pakistan are on the brink of an armed conflict, and the following is a comment on social media.*
>
> ⟨`comment`⟩
>
> *Do you find this content de-escalating? (Refer to the definition of de-escalating in the instructions)*
>
> **Q2:** *Do you think a Pakistani will also find this content de-escalating?*

Following standard practice [Weerasooriya *et al.*, 2023], a batch consists of 30 data instances for which we ask these two questions (first-person de-escalation, and vicarious de-escalation). We ran an initial pilot to ensure our survey ran smoothly.

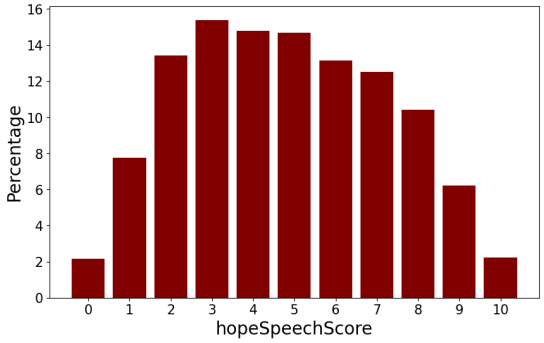

Figure 1: Data distribution with respect to *hopeSpeechScore*. The *hopeSpeechScore* of a given comment is the total number of raters who label the comment as *hope speech*. Overall, ten raters annotated each comment. A score of 10 implies all raters (across India and Pakistan) annotated the comment as *hope speech*. A score of 0 implies all raters are unanimous on the comment being *notHopeSpeech*.

Rater Demographics

Our overall dataset consists of 10,081 instances each of which is labeled by five Indian and Pakistani raters. Overall, this dataset is annotated by 1,639 unique Indian raters and 1,687 unique Pakistani raters. SI contains an in-depth review of the particulars regarding the rater demographics and rater compensation.

Disagreement Resolution

Prior literature has considered diverse approaches to resolving inter-annotator disagreements (e.g., majority voting [Davidson *et al.*, 2017; Wiegand *et al.*, 2019] or third objective instance [Gao and Huang, 2017]) or post-annotation adjudication [Breitfeller *et al.*, 2019]). In our work, raters are crowd workers whose identities are protected. Hence, no follow-on adjudication step is possible. For a given country, we consider the majority label. Since we consult five raters

for each instance in a binary labeling task, no tie-breaking is required.

4 Results and Discussion

4.1 RQ1: First Person De-escalation

RQ1: *How well do crowd workers align with experts and among themselves?*

Tables 2 and 3 summarize our results. We observe that both Indian and Pakistani annotators exhibit fair agreement with expert annotation. The majority labels from India and Pakistan also exhibit fair agreement (Cohen's κ 0.32). Gaps in annotation quality within experts and non-experts are well-documented in the literature [Hsueh *et al.*, 2009]. Hence, our observed disagreement among experts and crowd workers has precedence. Also, grounded in prior literature on efficient crowdsource task design [Finnerty *et al.*, 2013], we use a simplified definition of *de-escalating* and *not-de-escalating* content in our annotation study to reduce the cognitive load on the raters. This could account for some of the discrepancies between expert and non-expert annotations. That said, considering our task is subjective (and perhaps sensitive and contentious), our observed agreement is not out of line as reported in the literature as substantiated in what follows.

Demszky *et al.* [2020] released a dataset of 58K Reddit comments with fine-grained emotions categories. All raters in this study were from India. The observed Cohen's κ across the 27 emotion categories was 0.29 ± 0.14, a value comparable to the agreement (0.32) in our study. On a stance mining task, Khorramrouz *et al.* [2023] reported Cohen's κ ranging from 0.41 to 0.52. On a misogyny annotation task conducted by expert annotators, Guest *et al.* [2021] reported Fleiss' κ of 0.48. Sanguinetti *et al.* [2018] report category-wise $\kappa = 0.37$ for offence and $\kappa = 0.54$ for hate. Finally, our observed inter-country agreement is higher than Gomez *et al.* [2020] (κ = 0.15) and Fortuna and Nunes [2018] (κ = 0.17).

Let the *hopeSpeechScore* of a given comment be the total number of raters who label the comment as *hope speech*. Overall, 10 raters annotated each comment. A score of ten implies all raters (across India and Pakistan) unanimously annotated the comment as *hope speech*. A score of zero implies all raters are unanimous on the comment being *notHopeSpeech*. Figure 1 presents the distribution of *hopeSpeechScore* across our dataset. While many comments received a score between four to six indicating substantial disagreement, we also observe that 16.29% of the dataset had near-consensus labels (at least nine out of ten raters from both countries agreed on the label). This result underscores that bipartisanship exists; despite the palpable differences among the countries' political views and international policies, a substantial chunk of the dataset received near-consensus labels.

What makes a social web post contested across the two countries' raters? We first construct a subset, $\mathcal{D}_{unanimous}$ with comments with *hopeSpeechScore* 0, 1, 9, and 10 (1,642 instances) showing near-unanimous agreement. Next, we construct a subset, $\mathcal{D}_{contested}$, consisting of comments where both Indian and Pakistani raters are almost unanimous (four or more raters from one country), but with opposing labels (479 instances). These are comments that one country's raters strongly believe are de-escalating (or not de-escalating) while the other country's raters strongly disagree. For $\mathcal{D}_{unanimous}$ and $\mathcal{D}_{contested}$, we compute the respective unigram distributions $\mathcal{P}_{unanimous}$ and $\mathcal{P}_{contested}$. Next, for each token t, we compute the scores $\mathcal{P}_{unanimous}(t) - \mathcal{P}_{contested}(t)$, and $\mathcal{P}_{contested}(t) - \mathcal{P}_{unanimous}(t)$ and obtain the top tokens ranked by these scores (indicating increased usage in the respective sub-corpus). Table 4 lists the top tokens present in each sub-corpus. Manual inspection reveals that content that both countries unanimously agree on is of unequivocal calls for peace or contains dangerous speech. Table 5 lists a few illustrative examples with near-unanimous agreement among Pakistani and Indian raters. In contrast, strongly disagreed instances exhibited allegiance to the individual country's army and calls for revenge. Table 6 lists a few illustrative examples.

4.2 RQ2: Vicarious interactions

RQ 2: *How well does a Pakistani (an Indian) rater predict if a social media post will be deemed as de-escalating by an Indian (a Pakistani) rater?* Table 7 summarizes the vicarious de-escalation prediction of Indian and Pakistani raters. Recall that, for a given comment d, in this task, raters predict whether someone from the other country would classify d as *hope speech*. This study examines how well raters anticipate out-group *hope speech* labels. Now, we have information about say, what Pakistanis believe Indians find as de-escalating. By comparing this with first-person perspectives, we assess how well raters understand their nuclear adversary's viewpoint. Our results indicate that Indian and Pakistani raters can represent the values and opinions of citizens of their rival country to some extent.

4.3 Diverse Perspectives Matter

We have three sets of annotations: by experts, by Pakistani raters, and by Indian raters. Each of these sets are conducted independently of each other. For a given country, individual raters annotated independently as well. ***Do diverse perspectives lead towards data partitions with improved agreement?*** Prior literature has shown that selecting a subset of the dataset with a higher annotator agreement can help in modeling tasks [Jiang and de Marneffe, 2019b; Jiang and de Marneffe, 2019a]. In what follows, we demonstrate that conditioning the data on majority consensus among Indian and Pakistani raters creates a data partition with substantially improved agreement.

Let $Pak_{maj}(d)$ returns the label of document d obtained through majority vote among the Pakistani raters. Similarly, let $Ind_{maj}(d)$ returns the label of document d obtained through majority vote among the Indian raters. We define $\mathcal{D}_{consensus} \subseteq \mathcal{D}_{hope}$ as the set of documents where these two labels agree, i.e., $d \in \mathcal{D}_{hope}$ and $Pak_{maj}(d) = Ind_{maj}(d)$. We further define, $\mathcal{D}_{subjective} \subseteq \mathcal{D}_{hope}$ as the set of documents where these two labels disagree, i.e., $d \in \mathcal{D}_{hope}$ and $Pak_{maj}(d) \neq Ind_{maj}(d)$.

For all practical purposes, from the point of view of Indian raters (or Pakistani raters), $\mathcal{D}_{consensus}$ is an arbitrary partition of $\mathcal{D}_{hope}$. We can only construct that if we have access to both India and Pakistan's perspectives. Yet, when we recompute the agreement of vicarious de-escalation only on

		Pak	
		hopeSpeech	*notHopeSpeech*
Experts	*hopeSpeech*	81.30%	18.70%
	notHopeSpeech	32.87%	67.13%

(a) Cohen's κ is 0.37

		Ind	
		hopeSpeech	*notHopeSpeech*
Experts	*hopeSpeech*	87.17%	12.83%
	notHopeSpeech	37.16%	62.84%

(b) Cohen's κ is 0.37

Table 2: Confusion matrices between expert annotators and crowd workers from Pakistan (Table 2a) and India (Table 2b). Expert annotations are obtained from Palakodety *et al.* [2020a]. Each instance is annotated by five raters from India and five raters from Pakistan. For each country and a given instance, we use a majority vote to aggregate individual rater's verdicts.

		Ind	
		hopeSpeech	*notHopeSpeech*
Pak	*hopeSpeech*	63.37%	36.63%
	notHopeSpeech	26.06%	73.94%

Cohen's κ is 0.37

Table 3: Confusion matrices between crowd workers. Each instance is annotated by five raters from India and five raters from Pakistan. For each country and a given instance, we use a majority vote to aggregate individual rater's verdicts.

More presence in $\mathcal{D}_{unanimous}$	**More presence in $\mathcal{D}_{contested}$**
war, peace, pakistan, india, love, want, people, country, army, good, like, just, hate, hope, media, stop, solution, think, attack, respect	india, love, want, jai, pakistan, nuclear, hind, bharat, hero, good, army, weapons, abhinandan, think, years, respect, israel, brothers, right, peace, support

Table 4: Words with higher presence in $\mathcal{D}_{unanimous}$ (left) and $\mathcal{D}_{contested}$ (right).

$\mathcal{D}_{consensus}$ (shown in Table 8), we observe dramatic improvement in agreement. We hypothesize that there perhaps exists a bipartisan and more objective notion of de-escalation. Solely from annotation results by Indian raters (or Pakistani raters) it is impossible to estimate this bipartisan data partition. However, within this bipartisan data, both Indian and Pakistani raters exhibit substantially better ability to predict vicarious de-escalation. Not only that, the agreement score with expert annotations also improves (shown in Table 9) on this partition of the data.

We hypothesize that $\mathcal{D}_{consensus}$, where Indian and Pakistani majority labels agree, contains more objective instances, while $\mathcal{D}_{subjective}$ consists of cases where majority labels diverge. This divergence may stem from either (1) inherent subjectivity, making annotation difficult regardless of nationality, or (2) fundamentally opposing views between the two groups.

Word	**Example text and labels**
war	**Text:** *If war starts both countries have to face economic crisis and common public will face problems* **Label:** *hopeSpeech* (Pak); *hopeSpeech* (Ind)
	Text: *No war my dear countrymen. We want peace, peace, and peace.* **Label:** *hopeSpeech* (Pakistan); *hopeSpeech* (India)
solution	**Text:** *We must have a peaceful solution Or else condition will be critical. War isn't a long term solution.* **Label:** *hopeSpeech* (Pakistan); *hopeSpeech* (India)
	Text: *One solution, let's NUKE Pakistan and Mullahs all over the world.* **Label:** *notHopeSpeech* (Pakistan); *notHopeSpeech* (India)

Table 5: Illustrative example comments from $\mathcal{D}_{unanimous}$ with words that have more presence in $\mathcal{D}_{unanimous}$ as listed in Table 4.

Word	**Example text and labels**
jai (loosely translates to hail)	**Text:** *All the best IAF.. destroy the enemy kill like anything in Pakistan at terror camps,etc., Leaving to you. Vandemathram vandematram jai hind jai hind.* **Label:** *notHopeSpeech* (Pakistan); *hopeSpeech* (India)
revenge	**Text:** *India will take revenge for this just wait and watch.* **Label:** *notHopeSpeech* (Pakistan); *hopeSpeech* (India)
hero	**Text:** *Salute to all our great brave martyrs of pulwama...u r our real heros..* **Label:** *notHopeSpeech* (Pakistan); *hopeSpeech* (India)
army	**Text:** *Yes true God is with Pakistan army...the right people.* **Label:** *hopeSpeech* (Pakistan); *notHopeSpeech* (India)

Table 6: Illustrative example comments from $\mathcal{D}_{contested}$ with words that have more presence in $\mathcal{D}_{contested}$ as listed in Table 4.

Figure 2 supports the former, showing that $\mathcal{D}_{subjective}$ has considerably fewer unanimous in-group labels than $\mathcal{D}_{consensus}$. In what follows, we present a modeling experiment that shows that models trained on a specific country's perspective perform substantially poorly in the test data that overlaps with $\mathcal{D}_{subjective}$. This result indicates that it is perhaps the former – data instances in $\mathcal{D}_{subjective}$ are innately subjective and challenging to annotate regardless of nationality.

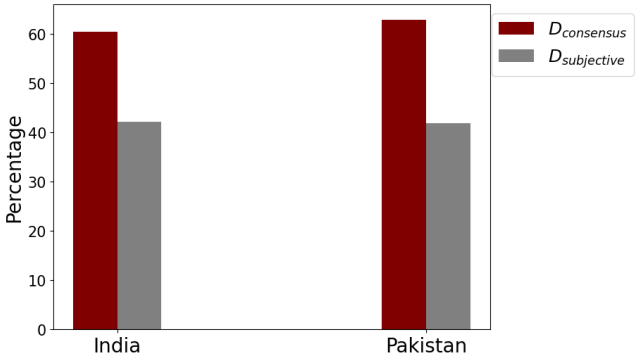

Figure 2: In-group unanimous agreement in $\mathcal{D}_{consensus}$ and $\mathcal{D}_{subjective}$ for India and Pakistan.

In supervised learning settings, do data instances from $\mathcal{D}_{subjective}$ and $\mathcal{D}_{consensus}$ behave differently? To answer this question, we train models solely on Indian ($\mathcal{M}_{Ind}$) and Pakistani ($\mathcal{M}_{Pak}$) majority labels. We create random 80:20 splits of $\mathcal{D}_{subjective}$ ($\mathcal{D}^{train}_{subjective}$ and $\mathcal{D}^{test}_{subjective}$) and $\mathcal{D}_{consensus}$ ($\mathcal{D}^{train}_{consensus}$ and $\mathcal{D}^{test}_{consensus}$) and construct train and test set with a proportional presence of $\mathcal{D}_{subjective}$ and $\mathcal{D}_{consensus}$. The train sets are further split into 90:10 for training and validation. We next fine-tune a high-performance LLM `Mistral 7B` [Jiang and others, 2023] and evaluate on the test which is essentially

Pak^{Ind}		Pak: *hopeSpeech*	Pak: *notHopeSpeech*
	hopeSpeech	83.43%	16.57%
	notHopeSpeech	46.52%	53.48%

(a) Cohen's κ is 0.38

Ind^{Pak}		Ind: *hopeSpeech*	Ind: *notHopeSpeech*
	hopeSpeech	80.81%	19.19%
	notHopeSpeech	45.30%	54.70%

(b) Cohen's κ is 0.36

Table 7: Confusion matrices between crowd workers on the vicarious perception of de-escalation computed on $\mathcal{D}_{hope}$. Each instance is annotated by five raters from India and five raters from Pakistan. For each country and a given instance, we use a majority vote to aggregate individual rater's verdicts. Pak^{Ind} represents vicarious de-escalation labels predicted by Indian raters for Pakistanis. Ind^{Pak} represents vicarious de-escalation labels predicted by Pakistani raters for Indians.

Pak^{Ind}		Pak: *hopeSpeech*	Pak: *notHopeSpeech*
	hopeSpeech	68.49%	31.51%
	notHopeSpeech	6.68%	93.32%

(a) Cohen's κ is 0.63

Ind^{Pak}		Ind: *hopeSpeech*	Ind: *notHopeSpeech*
	hopeSpeech	72.74%	27.26%
	notHopeSpeech	9.81%	90.19%

(b) Cohen's κ is 0.64

Table 8: Confusion matrices between crowd workers on the vicarious perception of de-escalation computed on $\mathcal{D}_{consensus}$. This result demonstrates a substantial increase in agreement as compared to Table 7. Pak^{Ind} represents vicarious de-escalation labels predicted by Indian raters for Pakistanis. Ind^{Pak} represents vicarious de-escalation labels predicted by Pakistani raters for Indians.

Experts		Crowd workers: *hopeSpeech*	Crowd workers: *notHopeSpeech*
	hopeSpeech	97.47%	2.53%
	notHopeSpeech	44.74%	55.26%

Cohen's κ is 0.55

Table 9: Confusion matrices between expert annotators and crowd workers from on $\mathcal{D}_{consensus}$. Expert annotations are obtained from Palakodety *et al.* [2020a]. This result demonstrates a substantial increase in agreement as compared to Table 2.

$\mathcal{D}^{test}_{consensus} \cup \mathcal{D}^{test}_{subjective}$.The test set is essentially $\mathcal{D}^{test}_{consensus} \cup \mathcal{D}^{test}_{subjective}$.

Table 10 summarizes our results. Our results indicate that even when a model is trained on data from a specific country's perspective and is tested on data labeled from the same country's perspective, struggles to correctly classify $\mathcal{D}^{test}_{subjective}$. Note that, estimating $\mathcal{D}_{subjective}$ is only possible if we consider diverse perspectives. Hence, our results indicate that consensus labels from both countries lead to data partitions that are objective and more learnable.

Model	Test Set	Accuracy	F1 Score
$\mathcal{M}_{Ind}$	$\mathcal{D}^{test}_{consensus}$	**0.87 ± 0.02**	**0.81 ± 0.03**
	$\mathcal{D}^{test}_{subjective}$	0.55 ± 0.04	0.53 ± 0.07
$\mathcal{M}_{Pak}$	$\mathcal{D}^{test}_{consensus}$	**0.88 ± 0.01**	**0.83 ± 0.01**
	$\mathcal{D}^{test}_{subjective}$	0.44 ± 0.02	0.43 ± 0.03

Table 10: Performance metrics of the fine-tuned `Mistral 7B` model. $\mathcal{M}_{Ind}$ is trained on majority labels from Indian raters. $\mathcal{M}_{Pak}$ is trained on majority labels from Pakistani raters. Each experiment is run on five random 80:20 splits of test-train data.

4.4 Modeling Out-group Preferences

Our final set of experiments investigate in-the-wild performance of models trained on our dataset. Grounded in behavioral economics [Kahneman *et al.*, 2021], *noise audit* measures outcome variability across multiple decision systems. In the web-toxicity literature, a resource bottleneck for in-the-wild assessment of offensive speech classifiers was the lack of large-scale annotated datasets. Weerasooriya *et al.* [2023] demonstrated that this requirement can be bypassed by conducting a noise audit on a large pool of unlabeled data. The intuition is that we do not need to know the ground truth to study outcome variability. Weerasooriya *et al.* [2023] trained a large number of content classifiers and ran inference on a vast pool of unlabeled instances and studied the disagreement across different content classifiers. In a similar vein, we conduct a noise audit between four models: one trained on Indian labels, one trained on Pakistani labels, and two others trained on labels from vicarious perspectives of India and Pakistan. We train four models: $\mathcal{M}_{Ind}$; $\mathcal{M}^{vicarious}_{Ind}$; $\mathcal{M}_{Pak}$; and $\mathcal{M}^{vicarious}_{Pak}$. $\mathcal{M}_{Ind}$ and $\mathcal{M}^{vicarious}_{Ind}$ are trained on the original Indian labels and the Indian labels provided by Pakistani annotators through the vicarious perspective, respectively. Similarly, $\mathcal{M}_{Pak}$ and $\mathcal{M}^{vicarious}_{Pak}$ are trained on the original Pakistani labels and Pakistani labels provided by Indian annotators through the vicarious perspective, respectively. These models are then evaluated on an unlabeled pool of 10,000 instances. We use `Mistral 7B-v0.3` for fine-tuning and inference.

Table 11 summarizes our results. We observe that the in-the-wild agreement across models trained on raters from one country and its vicarious perspective from the other country (i.e., $\langle \mathcal{M}_{Ind}, \mathcal{M}^{vicarious}_{Ind} \rangle$ or $\langle \mathcal{M}_{Pak}, \mathcal{M}^{vicarious}_{Pak} \rangle$) is greater than the agreement across models trained on annotations provided by raters from different countries. This result underscores that the differences in annotation across raters from different countries get carried forward to models trained on them. Furthermore, the high agreement between $\mathcal{M}_{Ind}$ and $\mathcal{M}^{vicarious}_{Ind}$ (0.83) and between $\mathcal{M}_{Pak}$ and $\mathcal{M}^{vicarious}_{Pak}$ (0.91) highlights the potential of leveraging vicarious perspectives as a reliable proxy for missing annotations when raters can represent the values of opinions of the out-group. When the

	$\mathcal{M}_{Ind}$	$\mathcal{M}_{Ind}^{vicarious}$	$\mathcal{M}_{Pak}$	$\mathcal{M}_{Pak}^{vicarious}$
$\mathcal{M}_{Ind}$	-	**0.834**	0.733	0.716
$\mathcal{M}_{Ind}^{vicarious}$	**0.834**	-	0.767	0.754
$\mathcal{M}_{Pak}$	0.733	0.767	-	**0.906**
$\mathcal{M}_{Pak}^{vicarious}$	0.716	0.754	**0.906**	-

Table 11: In-the-wild Cohen's κ agreement of models trained on original and vicarious perspective annotations provided by raters from different countries. A cell, $\langle i, j \rangle$, represents the Cohen's κ observed between $\mathcal{M}_i$ and $\mathcal{M}_j$ inferences run on an unlabeled pool of 10,000 instances. This result indicates that models trained on a given country's raters' first-person perception of de-escalation closely mimics models trained on the opposite country's raters' vicarious perception of de-escalation.

entities involved (e.g., raters from different countries) share cultural, linguistic, or contextual similarities, their vicarious perspectives capture inherent patterns of judgment and reasoning that align closely with direct annotations. This alignment demonstrates that models trained on vicarious annotations (e.g., $\mathcal{M}_{Ind}^{vicarious}$ or $\mathcal{M}_{Pak}^{vicarious}$) are effective at approximating models trained on direct annotations (e.g., $\mathcal{M}_{Ind}$ or $\mathcal{M}_{Pak}$), providing valuable insights. We finally conclude on a positive note with Table 12 that lists in-the-wild *hope speech* instances where all four models, both trained on first-person perception on de-escalation and vicarious perception on de-escalation, agree on.

Pakistan and India should have war against terrorism sort out kashmir problem otherwise the humanity will suffer and i love India and i from Pakistan
WAR BRINGS MUTUAL DISTRUCTION...OUR COUNTRY DONT HAVE WELL ESTABLISHED AIR DEFENSE SYSTEM....PLZ DONT FORGET IF WE GO FOR WAR YOUR FUTURE GENERATION WILL BE HAMPERED AND DEVELOPMENT WILL SLOW DOWN...AND ONE IMPORTANT THING YOUR TARGET IS TERRORISM...AND NOT THE INNOCENT CHILDREN OF PAKISTAN...WE DONT GET ANYTHING BY KILLING THOSE WOMEN AND CHILDREN WHO ARE INNOCENT ,,,,MATURE BRO....WE SHOULD MAKE PAKISTAN TO WORK AGAINST TERRORISM...
am neither Pakistani and neither Indian but I have spend time in both countries and I have friends in both countries. I definitely don't want war in both these beautiful countries with beautiful cultures. I respect both Imran khan and Sidhu for what they is doing to patch this senseless tiff between India and Pakistan that has been dragging down both these nations for far too long

Table 12: Random in-the-wild *hope speech* instances where all four models ($\mathcal{M}_{Ind}$, $\mathcal{M}_{Ind}^{vicarious}$, $\mathcal{M}_{Pak}$, and $\mathcal{M}_{Pak}^{vicarious}$) have consensus.

5 Discussions

We present a novel dataset on de-escalation amidst near-conflict scenarios with balanced participation from Indian and Pakistani raters. To our knowledge, this is the first annotation study of web-manifestation of de-escalation between nuclear adversaries conducted at a scale involving more than 1,000 raters from both countries (India and Pakistan). Our annotation study indicates that despite differences in political views and international policy, Indian and Pakistani raters show considerable bipartisanship in their perception of de-escalation, both first-person and vicarious. Our study further reveals that including diverse perspectives may aid in identifying a subset of data that is more objective and learnable. Finally, our experiments reveal that vicarious interactions could provide a viable path to train models empathetic to out-group values. Given social media's increasingly important role in understanding and analyzing modern conflicts [Zeitzoff, 2017], and that we are in the midst of two major ongoing wars, our study contributes to the timely and important topic of bipartisanship in de-escalation.

6 Limitations

Our study has the following limitations.

- *More raters per instance:* India is a large country with considerable cultural variation. Intra-country diversity of this proportion can hardly be captured within five annotators per instance. We hope our released dataset will open the gates for similar annotation studies on a larger scale. Also, in this study, we have not asked about raters' political positions. Prior literature indicates that raters' political leanings are associated with how they annotate offensive content [Sap *et al.*, 2022; Weerasooriya *et al.*, 2023]. We believe our research will open the gates for similar studies investigating the association of political leanings and *hope speech* annotation.
- *Demographic biases:* We observe that our study participants are predominantly young. It is unclear if this bias is due to the platform or the nature and requirements of our task. India and Pakistan have witnessed painful partitions and multiple wars with the goriest one happening in 1971. The sense of loss could be different among older people. Also, the current study has a skewed gender distribution. Studying intersectionality in datasets along diverse demographic factors and connecting that with the perceptions of de-escalation, both first-person and vicarious, can be a meaningful follow-on research direction.
- *The voice of the expatriates:* Both India and Pakistan have a strong diaspora presence in several other countries, particularly in the UK. Studying if continued exposure to a shared culture influences raters' ability to predict vicarious de-escalation could be interesting follow-on research.
- *Beyond English:* Future studies can move beyond social web expressions authored in English (e.g., in Hindi and Urdu) lending linguistic diversity to analyses.

Ethical Statement

We follow the widely accepted social media research ethics policies that allow researchers to use user data without explicit consent if anonymity is protected [Benton *et al.*, 2017]. Due to the subjective nature of the annotation, we expect some biases in the distribution of labels. Hence, any biases that may be found there are unintentional. Also, we do not collect any personally identifiable information from the annotators. Content moderation can be potentially gruesome and affect the mental health of the moderators [Solon, 2017]. We maintain a small batch size (30 YouTube comments). The annotated data we release include de-identified publicly available posts, where users understand public access and there is no expectation of privacy. Hence, we see no major ethical concern. We rather believe that this dataset will open the

gates for further research in the domain of web manifestation of modern conflicts.

Contribution Statement

Arka Dutta and Syed Mohhamad Sualeh Ali are equal-contribution first authors. Syed Mohammad Sualeh Ali conducted this research while on a Fulbright scholarship at the Rochester Institute of Technology.

Acknowledgments

The authors thank Shriphani Palakodety for sharing the dataset and helping us with the setup of the initial experiments. The authors thank Jaime G. Carbonell and Tom M. Mitchell for their valuable contributions to the *hope speech* literature.

References

[Ali, 1983] Tariq Ali. *Can Pakistan survive?: the death of a state*. Penguin Books London, 1983.

[Benesch *et al.*, 2016] Susan Benesch, Derek Ruths, Kelly P. Dillon, Haji Mohammad Saleem, and Lucas Wright. Counterspeech on Twitter: A field study. Dangerous Speech Project, 2016.

[Benesch, 2014] Susan Benesch. Countering dangerous speech: New ideas for genocide prevention. *Available at SSRN 3686876*, 2014.

[Benton *et al.*, 2017] Adrian Benton, Glen Coppersmith, and Mark Dredze. Ethical research protocols for social media health research. In *ACL EthNLP*, pages 94–102, 2017.

[Bose, 2009] Sumantra Bose. *Kashmir: Roots of conflict, paths to peace*. Harvard University Press, 2009.

[Breitfeller *et al.*, 2019] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *EMNLP-IJCNLP*, pages 1664–1674, 2019.

[Chandra *et al.*, 2021] Mohit Chandra, Manvith Reddy, Shradha Sehgal, Saurabh Gupta, Arun Balaji Buduru, and Ponnurangam Kumaraguru. " a virus has no religion": Analyzing islamophobia on twitter during the covid-19 outbreak. In *ACM HT*, pages 67–77, 2021.

[D'Anieri, 2023] Paul D'Anieri. *Ukraine and Russia*. Cambridge University Press, 2023.

[Davidson *et al.*, 2017] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *ICWSM*, volume 11, pages 512–515, 2017.

[Demszky *et al.*, 2020] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *ACL*, pages 4040–4054, 2020.

[Deng *et al.*, 2023] Naihao Deng, Xinliang Frederick Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. You are what you annotate: Towards better models through annotator representations. In *Findings of EMNLP*, pages 12475–12498, 2023.

[Dutta *et al.*, 2025] Sujan Dutta, Deepak Pandita, Tharindu Cyril Weerasooriya, Marcos Zampieri, Christopher M. Homan, and Ashiqur R. KhudaBukhsh. ARTICLE: annotator reliability through in-context learning. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence*, pages 14230–14237. AAAI Press, 2025.

[Finnerty *et al.*, 2013] Ailbhe Finnerty, Pavel Kucherbaev, Stefano Tranquillini, and Gregorio Convertino. Keep it simple: Reward and task design in crowdsourcing. In *SIG CHItaly*, pages 1–4, 2013.

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

[Gao and Huang, 2017] Lei Gao and Ruihong Huang. Detecting online hate speech using context aware models. In Ruslan Mitkov et al., editors, *RANLP*, pages 260–266, 2017.

[Gomez *et al.*, 2020] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications. In *WACV*, pages 1470–1478, 2020.

[Guest *et al.*, 2021] Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. An expert annotated dataset for the detection of online misogyny. In *EACL*, pages 1336–1350, 2021.

[Gupta *et al.*, 2023] Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md Shad Akhtar. Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation. In *ACL*, pages 5792–5809, 2023.

[Hande *et al.*, 2021] Adeep Hande, Ruba Priyadharshini, Anbukkarasi Sampath, Kingston Pal Thamburaj, Prabakaran Chandran, and Bharathi Raja Chakravarthi. Hope speech detection in under-resourced kannada language. *CoRR*, abs/2108.04616, 2021.

[Harrington *et al.*, 2019] Christina Harrington, Sheena Erete, and Anne Marie Piper. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *CSCW*, 3:1–25, 2019.

[Hengle *et al.*, 2024] Amey Hengle, Aswini Kumar, Sahajpreet Singh, Anil Bandhakavi, Md Shad Akhtar, and Tanmoy Chakroborty. Intent-conditioned and non-toxic counterspeech generation using multi-task instruction tuning with RLAIF. In *NAACL-HLT*, pages 6716–6733, 2024.

[Hsueh *et al.*, 2009] Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. Data quality from crowdsourcing: a study of annotation selection criteria. In *NAACL HLT SIGANN*, pages 27–35, 2009.

[Jiang and de Marneffe, 2019a] Nanjiang Jiang and Marie-Catherine de Marneffe. Do you know that florence is packed with visitors? evaluating state-of-the-art models of speaker commitment. In *ACL*, pages 4208–4213, 2019.

[Jiang and de Marneffe, 2019b] Nanjiang Jiang and Marie-Catherine de Marneffe. Evaluating bert for natural language inference: A case study on the commitmentbank. In *EMNLP-IJCNLP*, pages 6086–6091, 2019.

[Jiang and Marneffe, 2022] Nan-Jiang Jiang and Marie-Catherine de Marneffe. Investigating reasons for disagreement in natural language inference. *TACL*, 10:1357–1374, 2022.

[Jiang and others, 2023] Albert Q Jiang et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[Johnson and Goldwasser, 2018] Kristen Johnson and Dan Goldwasser. Classification of moral foundations in microblog political discourse. In *ACL*, pages 720–730, July 2018.

[Kahneman *et al.*, 2021] Daniel Kahneman, Olivier Sibony, and Cass R Sunstein. *Noise: A flaw in human judgment*. Little, Brown, 2021.

[Khorramrouz *et al.*, 2023] Adel Khorramrouz, Sujan Dutta, and Ashiqur R. KhudaBukhsh. For Women, Life, Freedom: A Participatory AI-Based Social Web Analysis of a Watershed Moment in Iran's Gender Struggles. In *IJCAI*, pages 6013–6021, 2023.

[KhudaBukhsh *et al.*, 2020] Ashiqur R. KhudaBukhsh, Shriphani Palakodety, and Jaime G. Carbonell. Harnessing code switching to transcend the linguistic barrier. In *IJCAI*, pages 4366–4374, 2020.

[Malik, 2002] Iffat Malik. *Kashmir: Ethnic conflict international dispute*. Oxford University Press Oxford, 2002.

[Mathew *et al.*, 2019] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. Thou shalt not hate: Countering online hate speech. In *ICWSM*, volume 13, pages 369–380, 2019.

[Mathew *et al.*, 2021] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *AAAI*, volume 35, pages 14867–14875, 2021.

[Mitchell, 1997] Tom M. Mitchell. *Machine learning, International Edition*. McGraw-Hill Series in Computer Science. McGraw-Hill, 1997.

[Nie *et al.*, 2020] Yixin Nie, Xiang Zhou, and Mohit Bansal. What can we learn from collective human opinions on natural language inference data? In *EMNLP*, pages 9131–9143, 2020.

[Ouyang *et al.*, 2022] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

[Palakodety *et al.*, 2020a] Shriphani Palakodety, Ashiqur R KhudaBukhsh, and Jaime G Carbonell. Hope speech detection: A computational analysis of the voice of peace. In *ECAI*, pages 1881–1889, 2020.

[Palakodety *et al.*, 2020b] Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. Voice for the Voiceless: Active Sampling to Detect Comments Supporting the Rohingyas. In *AAAI*, pages 454–462, 2020.

[Pandita *et al.*, 2024] Deepak Pandita, Tharindu Cyril Weerasooriya, Sujan Dutta, Sarah K Luger, Tharindu Ranasinghe, Ashiqur R KhudaBukhsh, Marcos Zampieri, and Christopher M Homan. Rater cohesion and quality from a vicarious perspective. In *Findings of EMNLP*, pages 5149–5162, 2024.

[Passonneau *et al.*, 2012] Rebecca J Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46:219–252, 2012.

[Pavlick and Kwiatkowski, 2019] Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *TACL*, 7:677–694, 2019.

[Pei and Jurgens, 2023] Jiaxin Pei and David Jurgens. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. *CoRR*, abs/2306.06826, 2023.

[PK *et al.*, 2021] Kumaresan PK, Chakravarthi BR, Cn S, García-Cumbreras MÁ, Jiménez-Zafra SM, García-Díaz JA, Valencia-García R, Hardalov M, Koychev I, Nakov P, and García-Baena D. Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In *LTEDI*, pages 61–72, 2021.

[Plank, 2022] Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *EMNLP*, pages 10671–10682, 2022.

[Rosenthal *et al.*, 2021] Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. SOLID: A large-scale semi-supervised dataset for offensive language identification. In *Findings of ACL*, 2021.

[Saha *et al.*, 2022] Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech. In *IJCAI*, pages 5157–5163, 2022.

[Samuel, 2023] MT Samuel. The israel-hamas war: Historical context and international law. *Middle East Policy*, 30(4):3–9, 2023.

[Sanguinetti *et al.*, 2018] Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Antonio Stranisci. An Italian Twitter corpus of hate speech against immigrants. In *LREC*, 2018.

[Sap *et al.*, 2022] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A

Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *NAACL-HLT*, pages 5884–5906, 2022.

[Sarkar *et al.*, 2020a] Rupak Sarkar, Sayantan Mahinder, and Ashiqur KhudaBukhsh. The non-native speaker aspect: Indian English in social media. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 61–70. Association for Computational Linguistics, 2020.

[Sarkar *et al.*, 2020b] Rupak Sarkar, Hirak Sarkar, Sayantan Mahinder, and Ashiqur R KhudaBukhsh. Social media attributions in the context of water crisis. In *EMNLP*, pages 1402–1412, 2020.

[Schofield, 2010] Victoria Schofield. *Kashmir in conflict: India, Pakistan and the unending war*. Bloomsbury Publishing, 2010.

[Solon, 2017] Olivia Solon. Facebook is hiring moderators. But is the job too gruesome to handle?, 2017. The Guardian.

[Staniland, 2013] Paul Staniland. Kashmir since 2003: Counterinsurgency and the paradox of "normalcy". *Asian Survey*, 53(5):931–957, 2013.

[Toon *et al.*, 2019] Owen B. Toon, Charles G. Bardeen, Alan Robock, Lili Xia, Hans Kristensen, Matthew McKinzie, Roy J. Peterson, Cheryl S. Harrison, Nicole S. Lovenduski, and Richard P. Turco. Rapidly expanding nuclear arsenals in pakistan and india portend regional and global catastrophe. *Science Advances*, 5(10):eaay5478, 2019.

[Tyagi *et al.*, 2020] Aman Tyagi, Anjalie Field, Priyank Lathwal, Yulia Tsvetkov, and Kathleen M Carley. A Computational Analysis of Polarization on Indian and Pakistani Social Media. In *SocInfo*, volume 12467, pages 364–379, 2020.

[Weerasooriya *et al.*, 2023] Tharindu Cyril Weerasooriya, Sujan Dutta, Tharindu Ranasinghe, Marcos Zampieri, Christopher M Homan, and Ashiqur R KhudaBukhsh. Vicarious offense and noise audit of offensive speech classifiers: Unifying human and machine disagreement on what is offensive. In *EMNLP*, pages 11648–11668, 2023.

[Wiegand *et al.*, 2019] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. Detection of abusive language: the problem of biased datasets. In *NAACL-HLT*, pages 602–608, 2019.

[Yoo *et al.*, 2021] Clay H Yoo, Shriphani Palakodety, Rupak Sarkar, and Ashiqur R KhudaBukhsh. Empathy and hope: Resource transfer to model inter-country social media dynamics. In *ACL NLP4PI*, pages 125–134, 2021.

[Zeitzoff, 2017] Thomas Zeitzoff. How social media is changing conflict. *Journal of Conflict Resolution*, 61(9):1970–1991, 2017.